UNIVERSITÀ DI PISA

UNIVERSITÀ DEGLI STUDI DI PISA

Dipartimento di Informatica
Corso di Laurea Magistrale in Informatica per l'economia e per l'azienda
(Business Informatics)

TESI DI LAUREA MAGISTRALE

SEGMENTAZIONE E MODELLAZIONE DEL COMPORTAMENTO DEGLI UTENTI DI UN PORTALE PER LA RICERCA DEL LAVORO

TUTORE ACCADEMICO
Prof. Mirco NANNI

TUTORE AZIENDALE
Dott. Raffaele SERRECCHIA

CANDIDATO
Marco ROSATO

ANNO ACCADEMICO 2014 - 2015

L'obiettivo di questa tesi è quello di studiare, da un punto di vista metodologico e analitico, il comportamento degli utenti che visitano il motore di ricerca **Jobrapido**, con lo scopo di studiarli e segmentarli attraverso la realizzazione delle varie fasi del processo KDD.
Partendo dai dati raccolti e memorizzati quotidianamente nei Data Warehouse di Jobrapido, la tesi si propone di studiare, comprendere e riconoscere i vari tipi di comportamenti che i diversi utenti assumono nel momento in cui utilizzano il motore di ricerca, per poi capire su quali leve operative agire per poter incrementare il loro ingaggio e migliorare la loro esperienza sul sito. Inoltre, i risultati ottenuti forniranno informazioni di supporto al management, che potrà così decidere meglio come accrescere il business dell'azienda.
L'intero progetto si basa sullo sviluppo di due tipi di analisi:

- la prima definisce i diversi tipi di profili, utilizzando tecniche di clustering per segmentare gli utenti in base ai loro comportamenti;
- la seconda mira a individuare determinati attributi che possono caratterizzare un profilo piuttosto che un altro, in modo da capire le eventuali differenze inter e intra clusters.

Nelle varie parti in cui si eseguono gli studi di Data Mining, i processi sono stati pianificati con attenzione, ma alcune decisioni sono state prese solo dopo aver avuto a disposizione e analizzato i risultati delle fasi precedenti. Quindi, il presente elaborato riporta anche le varie motivazioni che hanno spinto a compiere determinate scelte.

Nelle successive parti finali di ogni analisi vengono mostrati i risultati ritenuti più interessanti, attraverso report e visualizzazioni, in modo tale da fornire nuove indicazioni per i diversi scopi di business.

Indice

Elenco delle figure

Capitolo 1

INTRODUZIONE

1.1 Contesto generale

Carte di credito, transazioni bancarie, telefonini, acquisti in rete, social network, applicazioni dei computer e così via generano quotidianamente un volume enorme di **dati**. Sia che siano generati dall'utente della rete, sia che siano generati automaticamente da sistemi scientifici, questi dati rappresentano un **universo digitale** importantissimo e in costante crescita, prezioso per il business delle aziende che, spesso non sono del tutto consapevoli di possedere una ricchezza così vasta e soprattutto così a portata di mano.

La raccolta dei dati ha subito un notevole aumento grazie all'intensificarsi dei device in grado di automatizzare numerose operazioni, sia nel business sia nella vita privata: nasce così l'era dei ≪**Big Data**≫ la cui elaborazione, però, richiede strumenti, tecnologie e risorse che vanno ben di là dagli strumenti convenzionali di gestione e immagazzinamento dei dati. Risulta, dunque, fondamentale comprendere quali siano le reali opportunità e le implicazioni che possono scaturire dall'utilizzo costante di tecnologie in grado di raccogliere e analizzare i Big Data.

Essi sicuramente sono una grande opportunità in quanto aprono le porte verso nuovi modi di percepire il mondo e, conseguentemente, di prendere decisioni. Sono il nuovo strumento che rende ≪**misurabile**≫ la società e guidano verso una nuova scienza dei dati finalizzata a misurare la realtà in tutti i suoi aspetti.
Un'azienda che decide di analizzare e gestire in modo efficace i Big Data che la riguardano, dovrà implementare le tecnologie sostenendo dei costi, ma dovrà anche adottare una nuova cultura aziendale e utilizzare nuovi schemi mentali. In questo modo avrà la possibilità di mettere in pratica nuovi modelli di business, ricavandone dei vantaggi economici.
Ecco perché, a detta degli esperti, in quella che viene definita un'economia Data Driven, i Big Data potrebbero rappresentare il nuovo petrolio.

1.2 Presentazione del progetto e approccio adottato

Il motore di ricerca Jobrapido, come tantissime altre aziende, ha a disposizione una grande quantità di dati che quotidianamente vengono raccolti, analizzati e interpretati in funzione dei diversi modelli di business che l'azienda di volta in volta decide di realizzare. Infatti, come per ogni società di servizi online, il maggior guadagno si ottiene riuscendo a veicolare nel proprio portale il maggior numero di utenti possibili, aumentando il proprio traffico web e portando quindi a un incremento dei profitti.

Naturalmente, è possibile ottenere tutto questo solo se si riesce a offrire all'utente un prodotto di qualità, e se intorno all'azienda si costruisce una macchina efficiente fatta di persone e tecnologie che permettano entrambe di rispondere a tutte le esigenze degli utenti, migliorando costantemente la loro esperienza sul portale. Vista quindi l'importanza e il ruolo centrale che l'utente riveste in questo tipo di società, l'idea di base di questa tesi è stata quella di realizzare un primo approccio che permettesse di creare profili in base al tipo di utente che utilizza il portale, studiando il loro comportamento al fine di migliorare la loro esperienza sul sito, fornendo così informazioni di supporto alle decisioni per il management.
L'approccio adottato è suddiviso essenzialmente in due grandi tipi di analisi, utilizzando alcune tecniche di Data Mining e servendosi di alcuni algoritmi di Machine Learning.

La prima fase è stata quella di **segmentare** gli utenti, costruendo prima delle **metriche** ad hoc che hanno permesso di descrivere in modo puntuale e preciso il loro comportamento. Successivamente, partendo da queste metriche appena costruite, si sono applicati alcuni **algoritmi di cluster** per poter far emergere i diversi gruppi (profili) di utenti.

La seconda fase ha cercato di capire in modo più approfondito il comportamento dei diversi gruppi di utenti, cercando di individuare possibili **differenze** tra i vari profili creati. Questo tipo di analisi è risultata molto importante in quanto si voleva capire se i diversi utenti, appartenenti a profili diversi, avessero avuto un'esperienza sul sito diversa, portandoli quindi ad essere attivi o meno, sulla base delle azioni compiute o degli eventi accaduti. Partendo da questo presupposto, si sono costruite più di 35 metriche diverse che permettessero di descrivere in modo accurato il tipo di esperienza e le varie azioni che un utente avrebbe potuto compiere utilizzando il portale web; in questa seconda fase con alcuni algoritmi di classificazione è stato possibile individuare queste differenze.

Nella parte conclusiva, dopo aver interpretato i risultati, sono stati costruiti dei **set di regole di comportamento** che permettono di individuare con una buona accuratezza se un utente appartiene a un profilo piuttosto che un altro, in modo quasi previsionale, fornendo così informazioni necessarie per lo sviluppo di determinate features nel portale

per migliorare l'ingaggio e incrementare il numero di utenti sul sito.

1.3 Rassegna della letteratura

Nella realizzazione di questo progetto sono stati utilizzati alcuni testi teorici sul Data Mining e Machine Learning, alcune documentazioni per utilizzare al meglio gli strumenti scelti e diversi articoli relativi alla segmentazione e alla creazione di profili di web users.

Per la prima parte sono stati fondamentali alcune sezioni dei testi di [Pang-Ning Tan, Vipin Kumar 06], [Foster Provost 13], [Daniel T. Larose 12] e [Bing Liu 12].

Per quanto riguarda l'utilizzo dei software per poter effettuare tutte le analisi, sono state consultate le rispettive guide o direttamente dai siti web principali (per quanto riguarda Knime ed R), oppure utilizzando alcune guide dedicate per il software Weka [Frank, Hall 13]

Infine per tutta la restante parte di analisi si sono letti e interpretati alcuni articoli e casi di studio per poter aumentare la propria conoscenza su argomenti e progetti abbastanza simili. I testi consultati sono riportati per esteso nella bibliografia.

Trattandosi di una tesi di tipo sperimentale, l'approccio è stato studiato direttamente all'interno dell'azienda, e la letteratura utilizzata ha permesso di arricchire e migliorare il percorso e la struttura proposta.

1.4 Contenuto della tesi

Dopo il **primo capitolo** di introduzione relativo al mondo dei Big Data in cui è stato presentato il problema da affrontare e l'approccio scelto per portare a termine gli obiettivi previsti, la tesi si articola in altri 5 capitoli.
Ognuno di essi descrive e illustra, in modo puntuale e preciso, le varie fasi del progetto.

Nel **capitolo 2** vengono presentati i fondamenti teorici sui quali poggia l'attività di tesi, a partire da nozioni riguardanti il dominio del Data Mining, il processo KDD, fino agli algoritmi utilizzati nelle fasi di analisi come il Clustering e la Classificazione.

Il **capitolo 3** presenta l'azienda nel suo complesso, cercando di illustrare il funzionamento del portale e le diverse azioni che un utente può compiere in esso.

Nel **capitolo 4** viene descritta la prima fase di analisi, sviluppata utilizzando tecniche di cluster per poter iniziare a costruire dei profili sulla base del comportamento assunto dagli utenti. I risultati sono stati visualizzati e interpretati.

Il **capitolo 5** descrive la seconda fase di analisi, in cui sono state utilizzate tecniche di classificazione, per comprendere e prevedere il comportamento e le differenze degli utenti sulla base dei profili precedentemente individuati. Anche per questa seconda fase si è proceduto a visualizzare e interpretare i dati.

Il **capitolo 6** è relativo alla conclusione della tesi.

Capitolo 2

Background sul Data Mining e Knowledge Discovery

2.1 Definizione del termine

Il termine **Data Mining** indica l'insieme delle tecniche e delle metodologie che hanno per oggetto l'estrazione, a partire da una grande quantità di dati, di un sapere o di una conoscenza che altrimenti rimarrebbero sconosciuti e l'utilizzo di questa conoscenza per fini industriali o operativi.

In modo più specifico, il termine si riferisce all'applicazione di una o più tecniche che consentono l'esplorazione di grandi quantità di dati, con lo scopo di individuare le informazioni più significative e di renderle disponibili e direttamente utilizzabili per poter prendere delle decisioni.
L'estrazione di conoscenza, ossia delle informazioni significative, avviene, per esempio, tramite individuazione delle associazioni o di **pattern** nascosti nei dati. In questo contesto un pattern indica una struttura, un modello o, in generale, una rappresentazione sintetica dei dati.
Le aziende non vogliono più memorizzare semplicemente i dati, ma sentono l'esigenza, in modo sempre più crescente, di capire e interpretare questi dati, per poterli trasformare e ottenere informazioni preziose per il proprio business.
Quindi, a causa anche della rapida evoluzione del mercato che richiede una capacità di adattamento, il Data Mining permette di far fronte a questa esigenza. E' dunque importante riuscire a sfruttare la potenziale ricchezza di informazioni di cui si dispone per generare vantaggio competitivo.
Spesso il termine Data Mining viene utilizzato come sinonimo di *knowledge discovery in databases* (KDD), anche se sarebbe più preciso parlare di *knowledge discovery* quando ci si riferisce al processo di estrazione della conoscenza, e di Data Mining quando si intende una particolare fase del suddetto processo.

2.2 Perchè usare il Data Mining

Gli algoritmi di Data Mining sono stati sviluppati per far fronte all'esigenza di sfruttare il patrimonio informativo contenuto nei dati che vengono raccolti e di cui le aziende dispongono.
Le fasi di acquisizione dei dati, solitamente, non sono più un problema, in quanto ormai questi sono accessibili dalle miriadi di sorgenti Web disponibili o attraverso i Data Warehouse aziendali. Il problema, invece, è sapere estrarli, utilizzarli e trasformali in informazioni.

Molto spesso i dati si presentano in forma eterogenea, ridondante e non strutturata. Tutto ciò fa si che solo una piccola parte di essi venga analizzata. Considerando anche la rapida evoluzione del mercato che richiede capacità di adattamento, il saper sfruttare la potenziale ricchezza di queste informazioni che si hanno a disposizione, costituisce un enorme vantaggio. Per fare ciò è necessario disporre di strumenti potenti e flessibili.
La gran quantità di dati e la loro natura eterogenea, infatti, rende inadeguati gli strumenti tradizionali. Ecco alcune caratteristiche relative ai dati:

- **grande dimensione dei dati:** come già detto, le tecniche tradizionali di analisi risultano molto efficienti per insiemi di dati di piccoli dimensioni.
  Quando la quantità di informazioni da processare è notevole, la complessità di alcuni algoritmi aumenta in modo insostenibile e alcuni risultati che ne conseguono possono essere meno attendibili o interpretabili;

- **eterogeneità e complessità dei dati:** le tecniche di analisi tradizionale analizzano normalmente sorgenti dati di tipo omogeneo, quindi con attributi dello stesso tipo.
  Con l'innovazione tecnologica, con il moltiplicarsi dei dati e delle sorgenti da cui attingere per ottenere informazioni, questa omogeneità è piano piano scomparsa.
  Ormai si lavora con dati eterogenei, molto più complessi come i dati non strutturati, da cui è possibile ottenere informazioni preziose grazie alle tecniche di Data Mining;

- **multi-dimensionalità dei dati:** solitamente, si è abituati a lavorare su insiemi di dati contenenti centinaia o migliaia di attributi.
  Le tecniche di analisi dei dati tradizionali sono state sviluppate per insiemi a bassa dimensionalità e spesso, da un punto di vista computazionale, le prestazioni risultano essere peggiori nel momento in cui sono presenti molti attributi;

- **scalabilità:** con il crescere delle dimensioni, le tecniche di analisi classiche devono affrontare il problema legato alla loro scalabilità, superato dagli algoritmi di Data Mining.

I vantaggi del Data Mining si possono riassumere nei seguenti punti:

- trattamento di dati quantitativi, qualitativi, testuali, immagini e suoni;
- possibilità di elaborare un numero elevato di osservazioni;
- possibilità di elaborare un numero elevato di variabili;
- algoritmi ottimizzati per minimizzare il tempo di elaborazione;
- semplicità di interpretazione del risultato;
- possibilità di visualizzare i risultati.

2.3 Il processo KDD

Il termine *Knowledge Discovery in Databases* (KDD) deriva dal titolo di un workshop organizzato da Piatetsky-Shapiro[1] nell'ambito del Machine Learning (Detroit, 1989)

> "Knowledge discovery in databases is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data." [Fayyad, 1996]

Con questo termine ci si riferisce all'intero processo, interattivo ed iterativo, di scoperta della nuova conoscenza che consiste nell'identificazione di relazioni tra dati, che siano valide, nuove, potenzialmente utili e comprensibili mediante l'uso di machine learning, statistica, intelligenza artificiale e base di dati.Alcuni degli aspetti e delle caratteristiche più rilevanti del processo KDD sono:

- **linguaggio di alto livello**: la conoscenza scoperta è rappresentata in un linguaggio di alto livello, in modo tale che possa essere compresa dall'utente umano;
- **accuratezza**: ciò che si scopre deve rispecchiare accuratamente il contenuto della base dati. Misure di incertezza esprimono il livello di attendibilità della conoscenza estratta;
- **efficienza**: il processo di scoperta è efficiente. I tempi di risposta sono predicibili e accettabili.

La scoperta della conoscenza consiste in una sequenza iterativa di fasi, riportate nello specifico qui di seguito:

- **Data selection:** dopo un attento studio relativo ai fini del business vengono selezionati i dati ritenuti importanti su cui successivamente verranno effettuate le analisi;

[1] Gregory Piatetsky-Shapiro è presidente del KDnuggets, sito leader nel campo della Business Analytics, Big Data e Data Mining, ed è uno dei massimi esperti del settore.

- **Data cleaning:** è la fase relativa alla preparazione dei dati. Consiste nell'evidenziare gli aspetti più importanti, rimuovendo l'eventuale rumore, dati duplicati e gestendo la presenza di outliers e missing value;

- **Data integration:** in questa fase vengono integrati i dati acquisiti da più sorgenti diverse;

- **Data transformation:** i dati vengono trasformati e consolidati in formati appropriati per eseguire i diversi tipi di algoritmi e analisi;

- **Data Mining:** è la fase più importante, nella quale vengono scelti e utilizzati i diversi algoritmi a seconda degli obiettivi prefissati. Il risultato del processo di Data Mining è considerevolmente influenzato dalla correttezza delle fasi precedenti, come una fase di analisi esplorativa, atta all'individuazione dell'algoritmo di Data Mining e dei relativi parametri che garantiscono le performance migliori. Una volta individuate queste informazioni si possono eseguire gli algoritmi scelti, che restituiranno come output un insieme di pattern (informazione utile);

- **Pattern evaluation:** Interpretazione dei pattern trovati e possibile ritorno alle fasi iniziali per reiterare il processo con il fine di raffinarlo;

- **Knowledge presentation:** fase in cui sono presenti tecniche di visualizzazione e di rappresentazione della conoscenza, per presentare all'utente la conoscenza estratta dai dati.

Il processo KDD è articolato in diverse fasi che a livello generale raccolgono le azioni specificate in precedenza. Inoltre, prevede come dati in input dati grezzi e fornisce come output informazioni utili ottenute attraverso le fasi illustrate nella figura che segue.

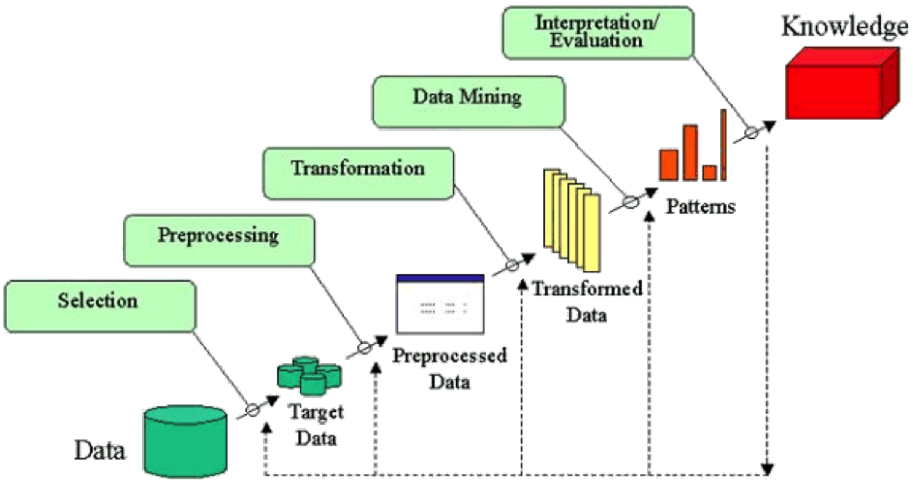

Figura 2.1: Le fasi del processo KDD

Selezione e data understanding (Data Selection)
I dati grezzi vengono segmentati e selezionati secondo alcuni criteri, al fine di ottenere un sottoinsieme di dati utili e rilevanti agli scopi dell'analisi e agli obiettivi stabiliti.
I database possono contenere diverse informazioni, ma il problema è considerare solo ciò che risulta importante ed essenziale.

Data quality (Data cleaning)
In questa fase si effettuano una serie di operazioni come la pulizia dei dati da eventuali rumori e da informazioni non corrette o non rilevanti, in modo tale da garantire qualità ed attendibilità ai dati che saranno successivamente oggetto di analisi.
Questo è un aspetto fondamentale in quanto non è possibile sperare di ottenere buoni risultati a partire da dati di scarsa qualità.

Trasformazione (Data transformation, Data integration)
Terminata la fase precedente, i dati, per essere utilizzabili, devono essere trasformati, cioè convertiti in determinati formati che saranno poi utilizzati dagli algoritmi di Data Mining; oppure se ne possono definirne di nuovi, attraverso l'uso di operazioni matematiche, statistiche e logiche sulle variabili.
Inoltre, soprattutto quando i dati provengono da fonti diverse, è necessaria una integrazione di questi dati, uniformandoli e omogenizzandoli, al fine di garantirne la loro consistenza.

Data Mining
Questa è la fase più importante, già illustrata in precedenza, in cui vengono applicati i vari algoritmi per l'analisi dei dati, scegliendo quelli più opportuni, a seguito delle analisi e degli studi fatti, in base ai propri scopi di business.
Il risultato sono dei pattern, ovvero dei modelli, che possono costituire un valido supporto alle decisioni.

Interpretazioni e valutazioni
Oltre che valutare questi modelli, è necessario capire anche quanto possono rivelarsi utili ai fini delle analisi. È dunque possibile, alla luce di risultati non perfettamente soddisfacenti, rivedere una o più fasi dell'intero processo KDD.
In generale, questa fase si occupa dell'interpretazione e valutazione dei risultati prodotti.

2.4 Il modello CRISP

Il Modello CRISP (Cross Industry Standard Process for Data Mining) è un prodotto neutrale, o meglio un progetto definito da un consorzio di numerose società per la standardizzazione del processo di Knowledge Discovery in Databases, pensato quindi per definire un approccio standard ai progetti di Data Mining.

Lo Scopo del progetto è quello di definire e convalidare uno schema d'approccio che sia indipendente dalla tipologia di business affrontato.
Dall'immagine si può notare come il ciclo di vita del processo sia formato da sei fasi la cui sequenza non è rigida, in quanto, in base alla qualità del risultato di ogni fase, è possibile ritornare alle fasi precedenti per poter ottenere un miglioramento o raffinamento nei risultati.

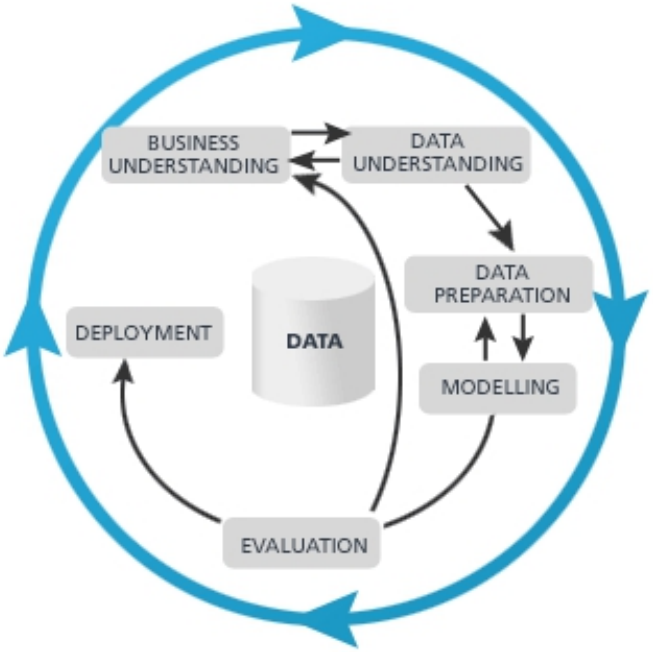

Figura 2.2: Descrizione del processo relativo al modello CRISP

Le sei fasi che compongono il modello sono:

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modelling
5. Evaluation
6. Deployment

Business Understanding: il primo passo è quello di conoscere in modo opportuno il settore di business in cui si opera. L'attenzione viene posta sugli obiettivi e requisiti del progetto da una prospettiva di business. Viene quindi definito anche il problema da risolvere.

Perciò, avendo chiare le idee sul business e sul problema che si vuole analizzare e risolvere, si procede alla conversione di questa conoscenza di settore nella stesura preliminare di un piano prefissato per raggiungere gli obiettivi stabiliti.

Data Understanding: una volta che sono stati definiti gli obiettivi da raggiungere, bisogna focalizzarsi sui dati, mezzo attraverso il quale sarà possibile ottenere i risultati attesi.
Quindi, questa fase prevede una iniziale raccolta dei dati e la successiva formulazione di ipotesi a seguito di quello che è emerso dalla visione preliminare. E' chiaro, inoltre, come le prime due fasi siano collegate tra loro dato che rappresentano l'individuazione dei fini e dei mezzi per un progetto di Data Mining.

Data Preparation: questa fase raccoglie tutte le operazioni che poi porteranno alla costruzione dell'insieme di dati finale, a partire dai dati grezzi, ai quali successivamente saranno applicate le tecniche di Data Mining. Questa fase comprende l'individuazione e la creazione di tabelle, attributi e variabili.
Inoltre, vengono applicate tutte le operazioni legate alla trasformazione e alla pulitura dei dati.

Modelling: Una volta che i dati sono pronti per essere analizzati, in questa fase vengono selezionate e applicate le varie tecniche e algoritmi che permettono di ricavare dei modelli da valutare.
Determinate tecniche, per poter essere applicate, necessitano di specifici formati o strutture rispetto alla forma dei dati, per cui è possibile che si debba ripetere la fase precedente per modificare il dataset iniziale e adattarlo alla tecnica di Data Mining specifica che si vuole utilizzare.

Evaluation: Prima di utilizzare il modello o i modelli costruiti, è necessario valutare il modello e i tutti i vari passaggi che si sono svolti per poterlo costruire, accertarandosi che attraverso tale modello sia possibile raggiungere gli obiettivi di business. Inoltre si ipotizza una futura applicazione del modello/i.

Deployment: questa è la fase finale che prevede l'utilizzo del modello o dei modelli creati e valutati per poter raggiungere gli obiettivi pianificati in partenza.

Le fasi appena descritte sembrano riprendere le fasi del processo di estrazione di conoscenza dai database (KDD). In realtà questo standardizzazione ingloba al suo interno le fasi del processo KDD e dimostra quanto già precedentemente spiegato riguardo il rapporto che sussiste tra Data Mining e KDD.

2.5 Clustering

Nato tra gli anni '60-'70, il **clustering** può essere definito come un insieme di tecniche di analisi multivariata dei dati, che mira alla selezione, individuazione e raggruppamento di elementi omogenei in un insieme di dati. Le varie tecniche di clustering si basano su concetti legati a delle misure di similarità tra gli elementi, concepite in termini di distanza in uno spazio multidimensionale.
In altre termini, tali algoritmi permettono di individuare gruppi omogenei di dati sulla base di variabili predefinite, in modo tale da massimizzare la distanza fra cluster e minimizzare le distanze intra-cluster.

L'analisi di clustering è stata largamente utilizzata in numerose applicazioni e gli obiettivi finali possono essere i più disparati, come il riconoscimento o l'isolamento di pattern caratteristici, l'analisi dei dati, lo studio degli effetti di diversi trattamenti sperimentali, la costruzione dei sistemi di classificazione automatica che consentono di immagazzinare informazioni e documenti, la ricerca di mercato e la riduzione della dimensione di grandi dataset.
Per esempio, in campo economico, il clustering può portare alla scoperta di gruppi distinti di clienti, caratterizzandoli in base ai loro acquisti. In campo statistico è possibile stratificare popolazioni da sottoporre a campionamento, mentre in biologia, le analisi di clustering possono essere utilizzate per categorizzare i geni con funzionalità simili, oppure per derivare le tassonomie delle piante e degli animali.
Tutto questo permette di ottenere informazioni aggiuntive sugli oggetti e sulle loro caratteristiche.

Esistono in letteratura un gran numero di algoritmi di clustering. Ovviamente, la scelta dell'algoritmo da utilizzare in un dato contesto dipende dalla tipologia di dati disponibili, dal particolare scopo dell'analisi e dall'applicazione stessa.
Per esempio, se le analisi di cluster vengono utilizzate con un fine descrittivo o esplorativo, è possibile testare diversi algoritmi sullo stesso dataset, per vedere i risultati generati da ciascuno di essi ed eventualmente confrontarli. In generale, i principali metodi di clustering possono essere così classificati:

- **Clustering partizionale:** Questi tipi di algoritmi generano gruppi di elementi in cui ogni oggetto appartiene esattamente a un solo cluster, in cui per definire l'appartenenza ad un gruppo viene utilizzata una distanza da un punto rappresentativo del cluster (centroide ecc...);

- **Clustering gerarchico:** Il risultato è un insieme di clusters nidificati che sono organizzati come un albero. Ciascun nodo (cluster) nell'albero (eccetto i nodi foglia) è l'unione dei suoi figli (sottoalberi) e la radice dell'albero è il cluster che contiene tutti gli oggetti;

- **Clustering basati sulla densità:** Questi algoritmi considerano i cluster come regioni dense di punti nello spazio dei dati, separando regioni a bassa densità dalle altre regioni a elevata densità.
  Questa tecnica viene utilizzata quando i cluster hanno forma irregolare o "attorcigliata", oppure in presenza di rumore o di outliers. Questi algoritmi possiedono quindi l'intrinseca capacita di rilevare cluster di forma arbitraria e di filtrare il rumore identificando gli outlier;

- **Esclusivo e non esclusivo:** In un clustering non esclusivo, i punti possono appartenere a più cluster. Viceversa, nei clusters esclusivi, ciascun oggetto è assegnato a un singolo cluster.

- **Fuzzy e non-fuzzy:** In un fuzzy clustering ogni punto appartiene a tutti i cluster con un peso di appartenenza che è tra 0 (assolutamente non appartiene) e 1 (assolutamente appartiene).
  In alcuni casi è possibile che venga imposto l'ulteriore vincolo che la somma dei pesi per ciascun oggetto deve essere uguale a 1;

- **Parziale e completo:** In un clustering parziale alcuni punti potrebbero non appartenere a nessuno dei cluster, mentre in un clustering completo ogni oggetto viene assegnato a un cluster.
  La motivazione per un clustering parziale è che alcuni oggetti in un data set possono non appartenere a gruppi ben definiti. Molte volte gli oggetti nel data set possono rappresentare rumori, outliers ecc..;

- **Eterogeneo e omogeneo:** In un cluster eterogeneo i cluster possono avere dimensioni, forme e densità molto diverse.

Nel caso specifico, si è inizialmente valutata la possibilità di applicare due diversi tipi di algoritmi, per confrontare successivamente i risultati ottenuti e capire quale potesse essere il migliore in termini di prestazioni e risultati.
Si sono utilizzati due tipi di algoritmi, il primo K-Means(partizionale) e il secondo DBSCAN(basato sulla densità) Si descrive brevemente il loro funzionamento per una migliore comprensione ed analisi.

2.5.1 K-means

L'algoritmo **K-Means** è una tecnica di clustering partizionante. Ogni cluster è caratterizzato da un centroide, e ogni punto viene assegnato al cluster che ha il centroide più vicino. Inizialmente l'algoritmo riceve in input un parametro K, che corrisponde al numero di cluster desiderati.

L'algoritmo segue una procedura iterativa: inizialmente crea K partizioni e assegna ad ogni partizione i punti d'ingresso o casualmente o usando alcune informazioni euristiche. Successivamente viene calcolato il centroide di ogni cluster. Ciascuno degli oggetti

rimanenti viene associato al cluster più vicino, in base alla distanza tra l'oggetto e la media del cluster. Il centroide di ciascun cluster è poi successivamente aggiornato in base ai punti assegnati al cluster. L'assegnamento e i passi di aggiornamento si ripetono finché l'algorimto non converge.

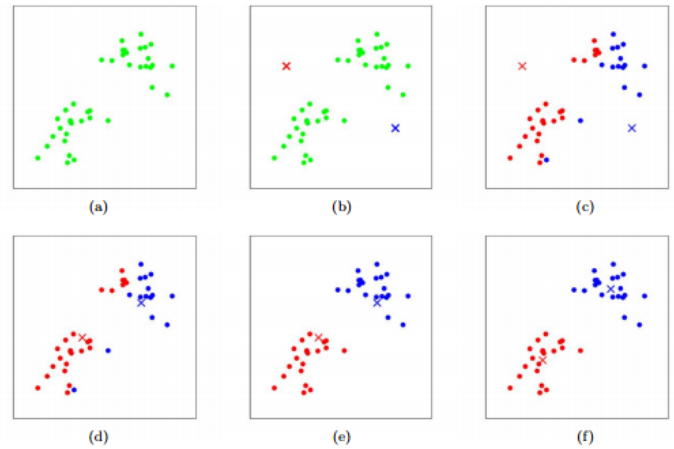

Figura 2.3: Sequenza di iterazione dell'algoritmo K-means

L'algoritmo k-Means presenta dei vantaggi ma anche degli svantaggi. I vantaggi sono:

- l'efficienza nel gestire grosse quantità di dati;
- la complessità computazionale dell'algoritmo è $O(tkmn)$ dove "m" è il numero di attributi, "n" il numero di oggetti, "k" il numero dei cluster e "t" è il numero di iterazioni sull'intero data set;
- spesso l'algoritmo termina con un ottimo locale ragionevolmente vicino a quello globale.

Gli svantaggi sono:

- i cluster hanno forma convessa, pertanto è difficile usare il k-means per trovare cluster di forma non convessa;
- il parametro k deve essere specificato a priori creando non pochi problemi nei casi in cui non si abbia alcuna informazione sui possibili modi in cui i dati possano dividersi;
- si potrebbe verificare la situazione in cui a uno o più cluster non vengono associati dei punti.
- è particolarmente sensibile agli outliers, che possono influenzare il modo netto il valore medio del cluster preso in considerazione.

2.5.2 DBSCAN

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) è il primo algoritmo che si basa sul concetto di densità. L'algoritmo costruisce ciascun cluster mettendo insieme regioni con densità sufficientemente alta, secondo determinati parametri. L'idea di base è che ogni punto appartenente ad un cluster debba avere nelle vicinanze, all'interno di un certo raggio, almeno un determinato numero di altri punti, ossia, la densità nelle vicinanze del punto considerato deve superare una certa soglia.

Bisogna specificare che l'algoritmo DBSCAN è una tecnica generale che può essere applicata a tutti i tipi di dati, cioè a pattern di qualunque natura, non solo punti.
L'algoritmo, in fase di impostazione, permette di definire due parametri:

- **Eps**: raggio dell'intorno di un punto;
- **MinPts**: numero minimo di punti nell'intorno.

Inoltre, sulla base dei valori precedentemente impostati, i vari punti vengono distinti dall'algoritmo in:

- **Core point:** sono i punti la cui densità è superiore a una soglia MinPts, perciò questi punti sono interni a un cluster;
- **Border point:** hanno una densità minore di MinPts, ma nelle loro vicinanze (ossia a distanza $<$ Eps) è presente un core point;
- **Noise point:** sono tutti i punti che non sono core point e border point.

L'algoritmo DBSCAN inizia selezionando un oggetto arbitrario p del dataset analizzato, calcolando il suo Eps: se al suo interno vi sono più di MinPts oggetti, p viene etichettato come punto appartenente a un cluster, altrimenti come rumore.
Successivamente si cerca di espandere il cluster appena trovato includendo iterativamente gli oggetti che rientrano nel raggio calcolato. Questo processo continua fino a quando il cluster non può più espandersi. Quindi si passa ad analizzare il successivo oggetto non visitato del dataset, fino a quando non si esauriscono tutti i punti.

Anche questo algoritmo presenta dei vantaggi e degli svantaggi. I vantaggi sono:

- è in grado di individuare clusters di forma arbitraria;
- non è sensibile al noise e agli outliers.

Gli svantaggi sono:

- è sensibile in modo considerevole nei confronti dei parametri Eps e MinPts, che a loro volta sono difficili da determinare;
- può introdurre pesanti costi di elaborazione di I/O arrivando fino a $O(n^2)$;
- può fondere erroneamente due cluster che siano congiunti da una stretta ma densa linea di punti (fenomeno del chaining).

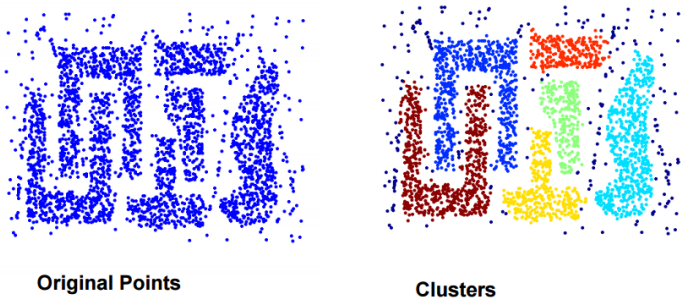

Figura 2.4: Esempio di come opera l'algoritmo DBSCAN

Di algoritmi di clustering, come illustrato precedentemente, ne esistono molti. Si è voluto approfondire il funzionamento di queste due tecniche, in quanto sono state quelle scelte per le analisi affrontate successivamente. Sono state applicate entrambe, per poi scegliere la tecnica migliore. Nel paragrafo successivo viene invece illustrato il funzionamento degli algorimti relativi agli alberi di classificazione, che saranno impiegati nella seconda fase di analisi legata alla segmentazione degli utenti.

2.6 Classificazione

La classificazione può essere definita come l'attività di apprendimento di una funzione target, che mappa ogni set di attributi a una delle classi predefinite, data una collezione di record caratterizzati da un set di attributi, di cui uno speciale denominato classe.
La funzione target, detta modello di classificazione, può essere utilizzata come modello descrittivo o predittivo. A differenza di altre tecniche, la classificazione richiede la conoscenza a priori dei dati e viene definita come una tecnica di apprendimento supervisionato.

Tipicamente il dataset di partenza viene suddiviso in due insiemi distinti chiamati **Training Set** e **Test Set**.
Il primo insieme, in cui la classe è nota a priori, permette di costruire il modello.
Il secondo insieme, invece, permette di validare il modello realizzato. Lo scopo di questa procedura è quello di valutare come si comporta il classificatore nei confronti di record di cui non si conosce la classe.

Gli algoritmi di classificazione portano all'identificazione di modelli che definiscono la classe cui appartiene un dato record con una determinata accuratezza (con il termine accuracy viene inteso il numero di record classificati correttamente dal modello, rispetto al totale dei record classificati.)
Le tecniche di classificazione possono essere utilizzate in numerosi contesti: per esempio, per quanto riguarda la ricerca, da parte delle banche, di categorie di clienti ai quali

concedere un prestito (in tal caso ogni record verrà etichettato come buon creditore o cattivo creditore e gli attributi possono essere per esempio dati dei clienti relativi all'età, al reddito, ecc..) o applicazioni di target marketing, con cui un'impresa può individuare, sulla base delle caratteristiche dei clienti presenti nel database, un proprio target di mercato allo scopo di rafforzare la propria posizione in un determinato settore (in tal caso etichettando ogni record del database come cliente fedele e cliente non fedele).

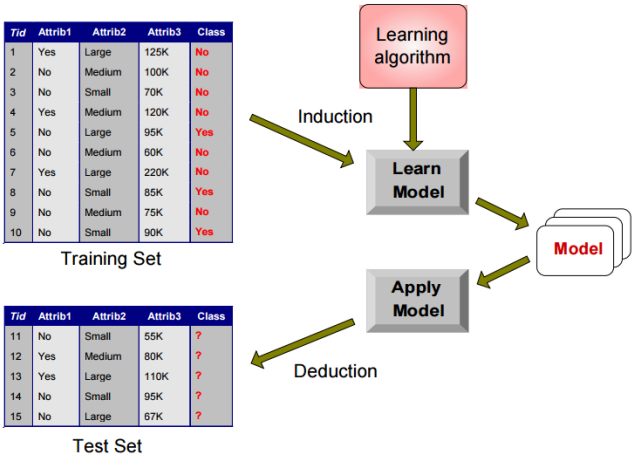

Figura 2.5: Il processo di classificazione

2.6.1 Alberi di decisione

Gli alberi di decisione costituiscono uno dei modi più semplice di classificare degli oggetti in un numero finito di classi. Inoltre, sono strutture molto conosciute nell'apprendimento supervisionato e sono strumenti molto utili per la risoluzione di svariati problemi reali.

La costruzione di modelli sulla base di alberi decisionali ha lo scopo di ripartire un dataset, attraverso una serie di passi. Questa fase avviene in base alle relazioni che legano la variabile target che si cerca di prevedere e una serie di variabili utilizzate come predittori. Il risultato del modello può essere rappresentato da un set di regole per ottenere i valori della variabile target o da una struttura ad albero.

Questi alberi vengono costruiti suddividendo ripetutamente i records in sottoinsiemi omogenei rispetto alla variabile target. La suddivisione produce una gerarchia ad albero, dove i sottoinsiemi (di records) vengono chiamati nodi e, quelli finali, foglie.
In particolare i nodi sono etichettati col nome degli attributi, gli archi (i rami dell'albero) sono etichettati con i possibili valori dell'attributo soprastante, mentre le foglie dell'albero sono etichettate con le differenti modalità dell'attributo classe che descrivono le classi di appartenenza.

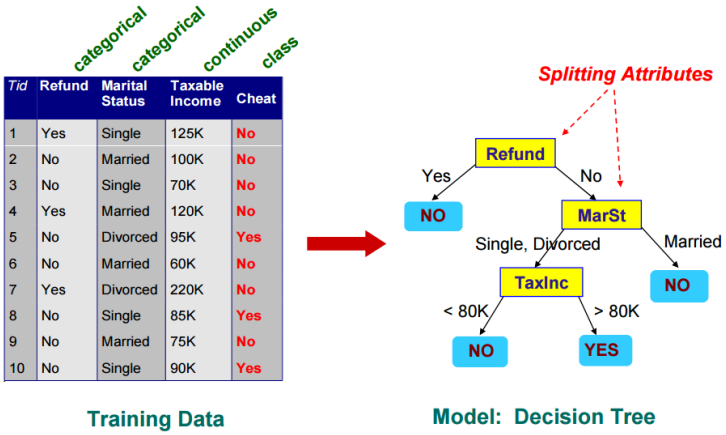

Figura 2.6: Un esempio di albero di decisione

Un oggetto viene classificato in base al percorso lungo l'albero che parte dalla radice fino alla foglia. I percorsi sono rappresentati dai rami dell'albero che forniscono una serie di regole. Quando un oggetto arriva in un nodo viene indirizzato su un ramo di uscita in base al risultato fornito dal test.
Il test, ovviamente, esamina i valori degli attributi dell'oggetto che sta analizzando. Infatti, ponendo un qualsiasi oggetto sulla radice dell'albero, questo può essere classificato, spostandolo nei rami opportuni in base alle regole precedentemente stabilite, fino a quando non si giunge ad una foglia. L'algoritmo base usato per costruire un albero decisionale è molto semplice ed efficiente:

1. si parte con un **singolo nodo** rappresentante tutti i record nel dataset;
2. si sceglie un **attributo** e si dividono i record in base al loro valore relativamente a questo attributo;
3. si ripete la **divisione** in tutti i nuovi nodi, fino a quando un criterio di stop non viene soddisfatto.

Questa procedura è conosciuta come **algoritmo di Hunt** e chiaramente esistono molti modi per implementarla. I tre aspetti molto importanti per capire il funzionamento degli algoritmi sono:

- come scegliere gli attributi di split;
- come determinare il migliore spit;
- quando fermarsi.

La costruzione degli alberi decisionali non viene fatta in modo casuale, ma viene realizzata rispettando determinati criteri al fine di ottimizzare il processo. Questi criteri permettono di capire quale sia la miglior selezione degli attributi che portano alla ramificazione dell'albero.

L'obiettivo di ogni algoritmo è quello di selezionare gli attributi più adatti per ottenere il migliore split per ogni nodo creato e determinare dei criteri di stop.
Per fare questo, i vari algoritmi esistenti differiscono principalmente per il tipo di misura di impurità che viene utilizzata per eseguire i migliori split nella creazione dell'albero. Le principali misure sono:

- Gini index
- Entropy
- Information Gain
- Gain Ratio
- Misclassification error

Gini index
Il gini index per un dato nodo t è uguale a:

$$Gini(t) = 1 - \sum_j [p(j|t)]^2$$

Per $p(j|t)$ si intende la frequenza relativa della classe j al nodo t. Il valore massimo si ottiene quando i records sono equamente distribuiti tra le classi, e corrisponde a $(1-1/n_c)$. Il valore minimo, invece, si ottiene quando tutti i records appartengono ad una classe, e corrisponde a 0.

Entropy
Intuitivamente, se tutti i record appartengono a una sola classe il nodo che risulterà avrà un alto valore di confidenza, cosa che invece non succede quanto i record sono equamente distribuiti tra tutte le classi.

Sia $p(c_i|n)$ la percentuale di record che appartengono alla classe c_i nel nodo n. L'entropia al nodo n è definita come:

$$Entropy(t) = -\sum_j p(j|t) \log p(j|t)$$

Il valore massimo si ottiene quando i records sono equamente distribuiti tra le classi, e corrisponde a $\log(n_c)$.
Il valore minimo, invece, si ottiene quando tutti i records appartengono ad una classe, e corrisponde a 0.

Information Gain

La formula relativa all'Information Gain è uguale a:

$$Gain_{split} = Entropy(p) - (\sum_{i=1}^{k} \left(\frac{n_i}{n}\right) Entropy(i))$$

Il nodo padre p viene suddiviso in k partizioni.
N_i è il numero di records nella partizione i.

Gain ratio

La formula è uguale a:

$$GainRATIO_{split} = \frac{Gain_{split}}{SplitINFO}$$

$$SplitINFO = -\sum_{i=1}^{k} \frac{n_i}{n} \log(\frac{n_i}{n})$$

Il nodo padre p viene suddiviso in k partizioni.
N_i è il numero di records nella partizione i.

Misclassification error

L'errore di classificazione ad un nodo t è uguale a:

$$Error(t) = 1 - maxP(i|t)$$

Il valore massimo si ottiene quando i records sono equamente distribuiti tra le classi, e corrisponde a $(1 - 1/n_c)$. Il valore minimo, invece, si ottiene quando tutti i records appartengono a una classe, e corrisponde a 0.

Una volta che sono state determinate le condizioni di split, rimane da capire quando l'algoritmo si ferma. Le condizioni sono le seguenti:

- un nodo non viene più suddiviso quando tutti i records appartengono a quel nodo;
- un nodo non viene più suddiviso quando tutti i records hanno valori similari su tutti gli attributi;
- si interrompe lo split quando il numero dei record nel nodo è inferiore a una certa soglia (data fragmentation).

Per concludere questa breve illustrazione del processo di classificazione, è necessario porre l'attenzione al problema di **underfitting** e **overfitting**, molto comuni quando vengono utilizzati questi tipi di algoritmi.

Per underfitting ci si riferisce ad un modello quando è troppo semplice e non consente una buona classificazione né del training set, né del test set.

Viceversa, per overfitting ci si riferisce ad un modello quando è troppo complesso, consente un'ottima classificazione del training set, ma una pessima classificazione del test set.
Il modello non riesce a generalizzare poiché è basato su peculiarità specifiche del training set che non si ritrovano nel test set.

Per quanto riguarda le metriche che permettono di valutare il modello, queste vengono elencate qui sotto:

- **Confusion Matrix** valuta la capacità di un classificatore sulla base dei seguenti indicatori:
  - TP (true positive): record correttamente classificati come classe Yes;
  - FN (false negative): record incorrettamente classificati come classe No;
  - FP (false positive): record incorrettamente classificati come classe Yes;
  - TN (true negative) record correttamente classificati come classe No.

  Se la classificazione utilizza n classi, la matrice di confusione sarà di dimensione $n \times n$.

- **Accuratezza** è la metrica maggiormente utilizzata per sintetizzare l'informazione di una confusion matrix.

  Accuratezza = $\frac{(TP+TN)}{(TP+TN+FP+FN)}$

  Equivalentemente potrebbe essere utilizzata la frequenza dell'errore

  Error rate = $\frac{(FP+FN)}{(TP+TN+FP+FN)}$;

- **Precision e Recall** sono due metriche utilizzate nelle applicazioni in cui la corretta classificazione dei record della classe positiva riveste una maggiore importanza.

  Recall misura la frazione di record positivi correttamente classificati. Valori elevati indicano che pochi record della classe negativa sono stati erroneamente classificati come positivi.

Precision, invece, misura la frazione di record risultati effettivamente positivi tra tutti quelli che erano stati classificati come tali. Valori elevati indicano che pochi record della classe positiva sono stati erroneamente classificati come negativi.

Recal = $\frac{TP}{(TP+FN)}$

Precision = $\frac{TP}{(TP+FP)}$;

- **F-measure** è una metrica che riassume precision e recall.

  F-measure = $\frac{(2*Recal*Precision)}{(Recal+Precision)}$

Capitolo 3

Il caso di Studio: Jobrapido

3.1 Jobrapido

Jobrapido è uno dei più grandi **motori di ricerca** (aggregatore) che analizza e raccoglie tutti gli annunci di lavoro presenti nel web in un'unica piattaforma, restituendo ai candidati una lista di offerte disponibili, classificandole e indicandone il grado di rilevanza.

Figura 3.1: Logo di Jobrapido

L'azienda permette quindi agli utenti di trovare un lavoro mediante l'utilizzo del web: attraverso la propria piattaforma è in grado di raccogliere tutti gli annunci di ricerca lavoro, separatamente inseriti in migliaia di siti web, nelle varie bacheche di lavoro, oppure diffusi dalle associazioni, e quelli esposti nelle pagine d'impiego delle singole società private, alla classica voce "lavora con noi" o similare.
Attraverso una serie di algoritmi, Jobrapido riordina tutte queste offerte di lavoro, che possono essere formulate in diverse decine di paesi esteri.

L'azienda è nata nel 2006, dall'idea del fondatore **Vito Lomele**, con l'intento di cambiare il modo di cercare lavoro. La sua esigenza nel trovare un lavoro lo ha spinto a creare un nuovo *aggregatore* per cercare di riunire tutte le offerte di lavoro, creando una piattaforma di job recruiting.
Nel 2006 l'azienda contava 10 persone, con un fatturato di 1 milione di euro. Durante la sua vita, l'azienda è stata acquisita diverse volte: la prima volta è stata comprata dalla Dgmt, il gruppo editoriale inglese che pubblica il Daily Mail, e successivamente dalla Symphony Technology Group, guidata dall'indiano Romesh T. Wadhwani.

Oggi si afferma come il motore di ricerca di annunci di lavoro più usato in Italia, presente in 58 paesi e che accoglie oltre 1,7 milioni di visitatori unici al mese e 25 milioni di visitatori unici nel mondo.

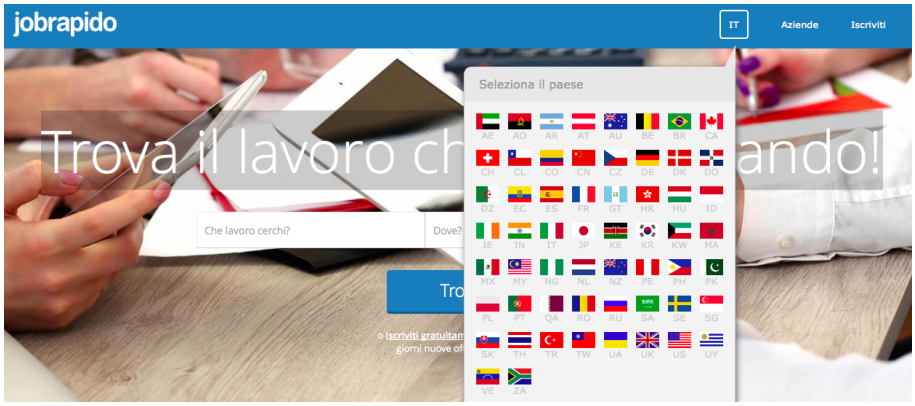

Figura 3.2: Jobrapido è presente in 58 paesi

3.2 Business Model

Per poter comprendere in modo approfondito il presente elaborato, si riporta di seguito una breve descrizione di come sia strutturata la parte di **background del portale**, e di come l'utente stesso (chiamato anche jobseeker) può agire e utilizzare la piattaforma di Jobrapido.
L'interfaccia del sito è molto semplice e intuitiva. Una volta che l'utente approda sul portale dovrà semplicemente compilare i tre campi seguenti:

- **tipo di lavoro:** in questo caso bisognerà inserire una o più parole chiave che permettono di descrivere la tipologia di lavoro ricercata;
- **dove:** la città, la regione, il luogo in cui si vuole andare a lavorare. (Jobrapido è presente in 58 paesi, quindi sarà anche possibile selezionare lo stato in cui ricercare il futuro impiego);
- **km:** è possibile specificare un raggio in km intorno alla città selezionata.

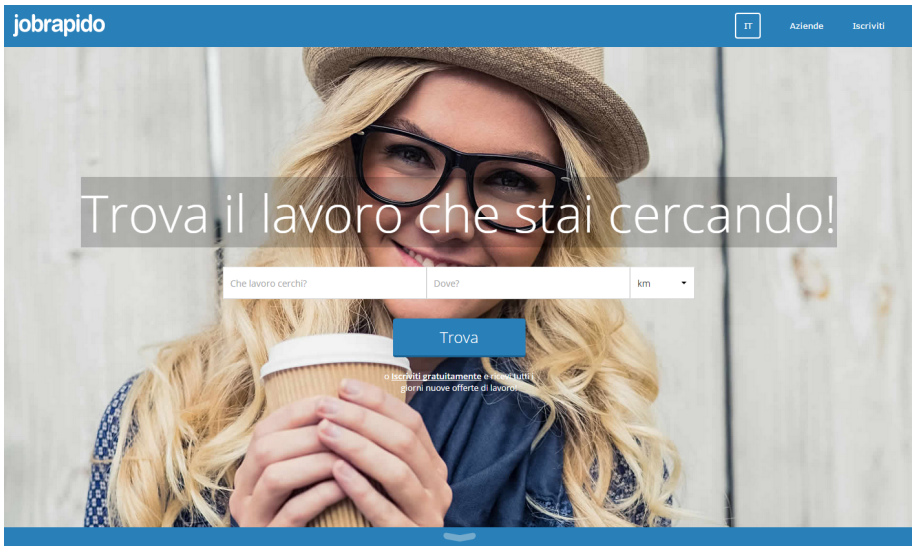

Figura 3.3: Home Page di Jobrapido

Una volta che i campi saranno stati compilati, Jobrapido mostrerà all'utente una serie di risultati, derivanti dalla ricerca effettuata.

Tramite i risultati mostrati, se gli utenti risultano interessati, con un semplice click verranno indirizzati al sito dal quale l'annuncio stesso ha avuto origine. Gli utenti possono così inoltrare direttamente il proprio curriculum o la domanda di lavoro.
Prima della visualizzazione, però, il portale darà la possibilità all'utente di decidere se iscriversi al servizio della **JobAllert**, oppure non iscriversi e navigare come un utente anonimo.

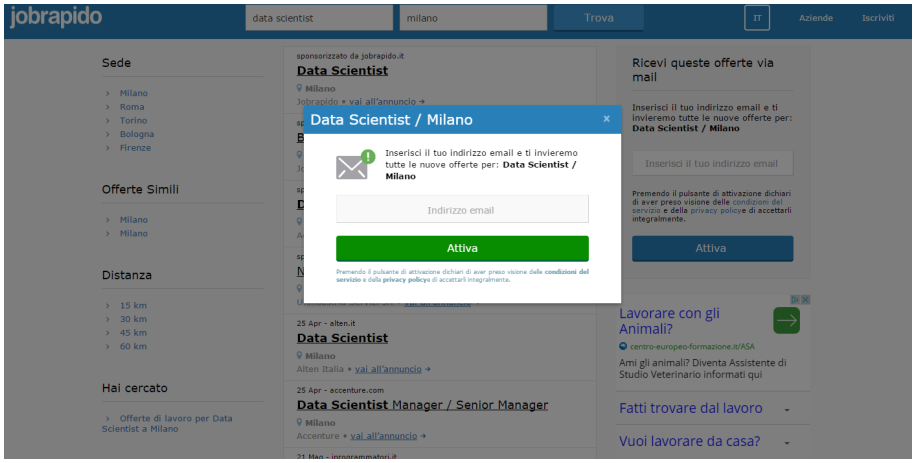

Figura 3.4: Popup per l'iscrizione alla Joballert di Jobrapido

L'iscrizione alla Joballert non richiede nessuna informazione personale, a eccezione della propria email.

Infatti, l'iscrizione permette all'utente di poter memorizzare fino a 10 ricerche diverse, in modo tale che il servizio possa inviargli eventuali proposte di lavoro che hanno un riscontro con gli statement (ricerche) salvati dall'utente e modificabili in qualsiasi momento.

A questo punto l'utente può iniziare la navigazione all'interno nel sito. Una volta compilati i campi, il sito mostrerà i risultati ottenuti.

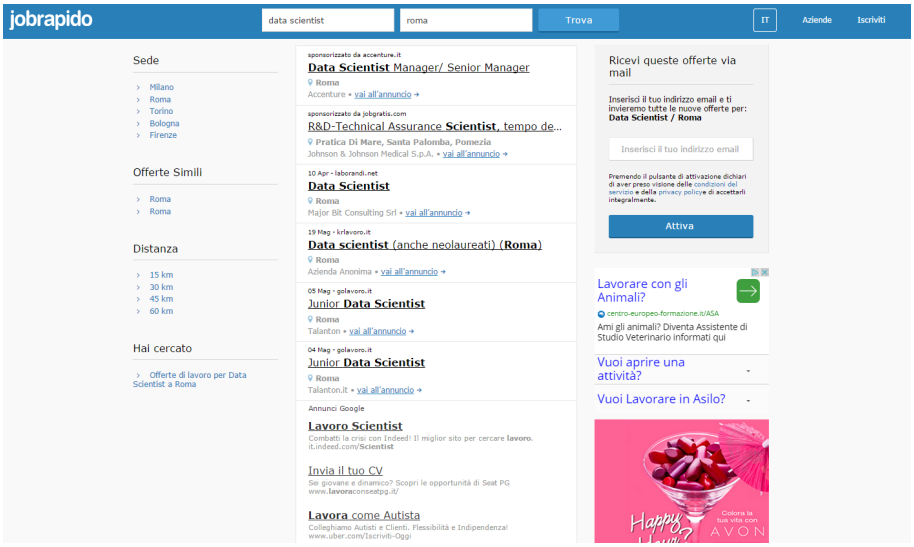

Figura 3.5: Pagine dei risultati di Jobrapido

All'utente verranno mostrati diversi risultati che si distinguono in:

- **job:** Sono i risultati che Jobrapido restituisce all'utente. Il ranking è deciso da Jobrapido stesso tramite calcolo della loro rilevanza. Possono essere:
  - sponsorizzati: sono quegli annunci per cui un cliente (di Jobrapido) ha pagato per poter comparire nei risultati di ricerca;
  - organici: sono quegli annunci gratuiti che vengono visualizzati perché sono pertinenti con i termini di ricerca di un utente.
- **contextual advertising :** Sono le Ad. Pubblicitarie, di cui Jobrapido non decide il contenuto, ma decide solo dove posizionarle all'interno del sito:
  - display: Ad. Pubblicitarie Visuali, dove dalla semplice visualizzazione si ha un guadagno;
  - text: Ad. Pubblicitarie testuali, dove il click genera un guadagno per Jobrapido stesso.

Durante la navigazione l'utente viene monitorato, generando una quantità di dati che verranno poi successivamente trasformate in informazioni e utilizzate per scopi di business e per masismizzarne l'esperienza.
L'origine dell'utente, intendendo il meccanismo mediante il quale ha permesso al jobseeker di giungere al sito, è molto importante.
Infatti, le visite provenienti da un anonimo jobseeker possono essere di tipo:

- **Brand direct:** La visita deriva da una conoscenza di Jobrapido da parte dell'utente (per esempio, immissione dell'URL oppure ricerca di Jobrapido direttamente sulla barra di ricerca di Google);

- **Paid:** Visite comprate (AdWords). L'utente ha cliccato un annuncio generato da AdWords;
- **Free oraganic:** L'utente è giunto sul portale tramite i risultati organici generati dai motori di ricerca in modo gratuito.

Questa è una breve e semplice descrizione delle azioni e del percorso che l'utente compie durante la navigazione sul sito di Jobrapido.

Naturalmente l'utente può rimanere anonimo nella sua navigazione, ma ciò non impedisce di non monitorarlo, in quanto ad ogni jobseeker viene associato un codice univoco la prima volta che entra nel sito.
Successivamente questo verrà riconosciuto durante le eventuali future visite, a meno che non succeda qualcosa (tipo cambio di dispositivo) che non permetta più il suo riconoscimento e tracciamento, e quindi verrà etichettato come nuovo utente.

Il discorso cambia se l'utente si è sottoscritto e/o confermato. Una volta che l'utente immette l'email nel pannello di iscrizione, l'utente è sottoscritto. Se poi l'utente conferma l'iscrizione dalla propria casella di posta elettronica, allora sarà confermato.
In questi casi per poterlo riconoscere, oltre al numero identificativo, ci sarà il proprio indirizzo email, anch'esso univoco.

3.3 Data Model

Tutti i dati raccolti giornalmente da Jobrapido vengono costantemente memorizzati nei rispettivi **Data Warehouse**, monitorati e analizzati, in base al tipo di esigenza richiesta.
Per Data Warehouse si intende una raccolta di dati integrata, orientata al soggetto, variabile nel tempo e non volatile di supporto ai processi decisionali, distinguendosi dai classici data base per alcune caratteristiche quali:

- **orientato ai soggetti di interesse:** è orientato a temi aziendali specifici, quindi considera i dati di interesse ai soggetti dell'organizzazione e non quelli rilevanti ai processi organizzativi. In un data Warehouse i dati vengono archiviati in modo da essere facilmente letti o elaborati dagli utenti. L'obiettivo è quello di fornire dati organizzati in modo tale da favorire la produzione d'informazioni;

- **integrato:** requisito fondamentale di un Data Warehouse è l'integrazione dei dati raccolti, in quanto c'è necessità di dare coerenza ai dati provenienti da diverse applicazioni, progettate per scopi diversi. Poiché i manager, per poter prendere le loro decisioni necessitano di ogni possibile fonte di dati, relativa (o meno) all'azienda, il problema da affrontare è quello di rendere questi dati accessibili e omogenei in un unico ambiente. Di tutti questi tipi di dati, il DW restituisce una visione unificata;

- **rappresentativo dell'evoluzione temporale:** i dati archiviati all'interno di un Data Warehouse coprono un orizzonte temporale molto più esteso rispetto a quelli archiviati in un database. Nel Data Warehouse sono contenute una serie d'informazioni relative alle aree d'interesse che colgono la situazione relativa a un determinato fenomeno in un determinato intervallo temporale piuttosto esteso. Quindi sono contenuti i dati che sono precedenti da 5 a 10 anni, o più vecchi, che devono essere usati per confronti, tendenze e previsioni. Questi dati non possono essere modificati;

- **non volatile:** questa caratteristica indica la non modificabilità dei dati contenuti nel data Warehouse, che consente accessi in sola lettura. Perciò i dati non possono essere modificati o cambiati in nessun modo una volta che sono entrati nel data warehouse, ma possono essere solo caricati e consultati.

Per motivi di riservatezza non è stato possibile riportare una struttura completa del Data Model di Jobrapido. Ma è possibile, invece, descrivere le motivazioni che hanno portato all'utilizzo dei data warehouse *HP Vertica*.

Jobrapido ha la necessità di gestire 10 TB di dati storicizzati, con una media di 50 GB di nuovi dati che si aggiungono giornalmente. Inoltre, operando in 58 paesi diversi, gestisce 12 fusi orari. Quindi, aveva la necessità di utilizzare un sistema che fosse:

- **scalabile:** l'infrastruttura doveva essere in grado di crescere facilmente all'aumentare delle dimensioni;
- **flessibile:** doveva essere facile arricchire il modello dati, generare nuovi report con dati esistenti senza dover creare nuovi datamart;
- **solido:** l'infrastruttura doveva possedere un disaster recovery efficiente. Per disaster recovery si intende l'insieme delle misure tecnologiche e logistico/organizzative atte a ripristinare sistemi, dati e infrastrutture necessarie all'erogazione di servizi di business a fronte di gravi emergenze che ne intacchino la regolare attività;
- **performante:** Era necessario possedere un'infrastruttura su cui si potessero effettuare delle interrogazioni con estrema efficienza e velocità.

Dopo alcuni studi e confronti, l'azienda ha deciso di utilizzare il data warehouse creato da HP, in quanto rispondeva a tutte queste esigenze in modo superiore rispetto ad altri data warehouse.

Capitolo 4

Analisi di Clustering

4.1 Descrizione del procedimento

Per poter migliorare ed essere sempre più competitiva sul mercato, un'azienda deve analizzare con attenzione tutte le innumerevoli informazioni in suo possesso.
Nello specifico di Jobrapido, l'analisi del comportamento degli utenti all'interno del portale, attraverso la loro segmentazione, risulta importante per capire quali siano le caratteristiche che spingono un utente a tornare più volte sul sito, o a utilizzare maggiormente tutte le funzionalità offerte dal servizio; in questo modo l'azienda può individuare le *features* su cui deve puntare per soddisfare le richieste dei propri clienti e raggiungere gli obiettivi che si è prefissata. In questo capitolo e nei capitoli successivi verrà descritto il processo, strutturato in diverse parti distinte, che ha portato ad analizzare il comportamento di alcuni utenti, tentando di crearne dei profili specifici e indagando sulle varie differenze per poter capire dove far leva per aumentare il guadagno che l'azienda può ottenere dai singoli utenti.

Nella prima fase del progetto, si è implementato un processo basato sull'utilizzo di tecniche di **clustering** con due scopi:

- **segmentare gli utenti** sulla base di alcune caratteristiche, utilizzando metriche costruite ad hoc;
- **creare profili** di utenti caratterizzati da comportamenti simili al loro interno e dissimili tra i profili stessi.

Questa prima fase ha permesso di etichettare gli utenti in base al loro cluster di appartenenza, in modo tale che, nella seconda fase del progetto, potessero essere applicati alcuni algoritmi di classificazione supervisionata che consentissero di studiare le differenze tra i vari cluster e di individuare eventuali attributi in grado di spiegare cosa ha portato un jobseeker a far parte di un gruppo piuttosto che di un altro.

Tutte queste informazioni sono utili al management per poter eventualmente prendere delle decisioni operative e sfruttare determinate leve per migliorare l'esperienza sul sito da parte dell'utente e quindi ottenere un maggior guadagno da essi. Più un utente torna sul sito, più un utente utilizza il sito, più l'azienda ha la possibilità di ottenere una remunerazione dall'utente stesso. Per questo motivo, capire quali siano le caratteristiche che possono spingere un utente a tornare più spesso e a utilizzare sempre di più i servizi offerti, può essere un ottimo punto di partenza per elaborare delle nuove strategie di mercato.

Questo capitolo è così strutturato:

- illustrazione dei **dataset** e dei tipi di dati elaborati e utilizzati per effettuare le varie analisi;
- spiegazione del procedimento che ha portato alla creazione delle **metriche** utilizzate nella fase principale di clustering;
- commento dei **risultati** ottenuti con le rispettive visualizzazioni e interpretazioni.

4.1.1 Scelta del dataset

La scelta del dataset da utilizzare per le analisi è stata molto elaborata, in quanto Jobrapido è presente in 58 paesi e il numero di utenti attivi è enorme.
Il primo passo, quindi, è stato quello di analizzare con attenzione la situazione e capire sia lo **stato** (denominato country) da cui estrarre i dati, sia **l'intervallo di tempo** su cui effettuare le analisi.
La scelta è ricaduta su 4 country: Australia, Francia, Regno Unito e Italia.

Le motivazioni sono state le seguenti:

- i quattro stati selezionati fanno parte di quelli in cui Jobrapido è maggiormente utilizzato: questo consente di avere a disposizione dati numerosi e precisi;
- in questi stati c'era la certezza che in determinati mesi dell'anno non fosse stata apportata al portale alcun tipo di modifica o di inserimento di nuove *features*. Questo ragionamento è stato prezioso in quanto per analizzare il comportamento degli utenti era necessario studiare un periodo temporale in cui si avesse la certezza che ogni utente avesse visto le stesse cose e avesse utilizzato gli stessi servizi, e che quindi avesse avuto lo stesso tipo di esperienza. Ciò che l'utente poteva fare sul portale doveva essere uguale per tutti gli utenti analizzati.

Per semplicità di illustrazione nel progetto si riportano i risultati e le visualizzazioni ottenute relativamente solo alla country **Italia**, sulla quale ci si è soffermati maggiormente. Non sono stati riportati i risultati delle altre country analizzate in quanto i valori e le informazioni non presentavano differenze rilevanti.

L'Italia è stata la country di analisi principale e lo scopo era mostrare il procedimento che aveva portato a ottenere i risultati mostrati in seguito. Anche la scelta dell'**intervallo temporale** è stata molto ponderata, per diversi motivi:

- bisognava scegliere un periodo temporale da cui fosse possibile estrarre un campione abbastanza numeroso per poter effettuare delle analisi.
- l'intervallo temporale doveva essere riferito a un periodo in cui all'interno del sito non fosse stata apportata alcuna modifica, in modo tale che tutti gli utenti analizzati avessero avuto lo stesso tipo di esperienza.

Dopo tutte queste considerazioni, si è deciso di esaminare tutti gli utenti sottoscritti e confermati al portale di Jobrapido tra il 1 gennaio 2015 e il 31 gennaio 2015, per poi considerare solo le visite effettuate dagli utenti dal momento della loro conferma fino al 31 Marzo 2015.
Il periodo preso a campione è di 3 mesi, con la variante che gli utenti analizzati sono solo quelli che si sono sottoscritti e confermati nel primo mese, in modo tale da garantire in modo maggiore una medesima esperienza dell'utente sul sito e quindi lavorare su un campione il più possibile omogeneo. Il dataset esaminato per la country Italia conteneva esattamente un campione di 92.634 utenti.

La sottoscrizione e la conferma di un utente ha permesso di tenere traccia tramite il rispettivo indirizzo email, e non tramite il numero identificativo, che veniva assegnato da Jobrapido stesso. Infatti, uno stesso utente entrando dal proprio personal computer e successivamente da un dispositivo mobile, avrebbe avuto due numeri identificativi diversi, e quindi si sarebbe perso questo tipo di informazione. Al contrario, analizzando solo gli utenti confermati, la chiave primaria sarebbe stato l'indirizzo email, e quindi si sarebbero avute informazioni molto più precise.

I dati analizzati, sono quelli che vengono raccolti quotidianamente da Jobrapido e si riferiscono alle operazioni che ogni jobseeker compie all'interno del portale, a partire dalle ricerche effettuate, il numero di risultati che vengono visualizzati a seguito della ricerca, il tipo di dispositivo e layout, quante pagine vengono visionate, quanto dura la singola visita, quante visite sono state effettuate, quali tipi di annunci e di siti sono stati visualizzati, quali tipi di pubblicità, quanti click sono stati fatti ecc. Sono questi i tipi di dati che si sono successivamente utilizzati e aggregati per poter creare le metriche necessarie all'analisi di clustering.

4.1.2 Tipologia di metriche

L'idea di partenza è stata quella di costruire delle metriche con i dati a disposizione, le quali sarebbero state utilizzate successivamente come attributi per il processo di clustering. Bisognava delineare un contesto accurato e preciso e perciò era necessario capire in che modo era possibile descrivere il comportamento degli utenti sul sito sulla base dei

dati messi a disposizione e, quindi, su quali basi si sarebbero costruiti i vari profili dei jobseekers. Perciò si è deciso di concentrarsi su due aspetti principali che coinvolgono il comportamento degli utenti: la loro **frequenza di accesso** e il loro **tasso di utilizzo del portale**.

Per frequenza d'accesso si intende la **quantità di accessi** effettuati al sito internet, vale a dire quanto spesso un jobseeker visita il portale.
Per utilizzo si intende il **modo** in cui il jobseeker si è servito delle funzionalità offerte da Jobrapido, cioè quante operazioni sono state effettuate.

L'idea era quella di cercare di descrivere e capire che tipo di utenti accedevano al portale: per esempio, possono esserci gruppi di utenti che visitano molto spesso il sito ma non fanno alcun tipo di operazione rilevante, oppure possono esistere gruppi di utenti che visitano molto poco il portale, ma in quelle rare occasioni utilizzano a pieno i servizi offerti, aumentando il loro tasso di utilizzo.

Una volta definite le aree di interesse, era necessario costruire delle metriche appropriate che fossero in grado di rappresentare con dei valori precisi sia la frequenza e sia l'utilizzo del portale. Questo avrebbe poi permesso di etichettare ogni utente in svariati profili, utilizzati nella seconda parte del progetto per indagare sulle possibili differenze tra di loro. Nei paragrafi successivi verrà illustrato il procedimento di data quality per la costruzione delle metriche nelle varie fasi che compongono l'analisi di clustering e il processo che porta alla determinazione dei risultati.

4.1.3 Software e tools

L'estrazione dei dati è avvenuta mediante il software **SAP B.O.**, definito come un portafoglio di strumenti e applicazioni perfettamente integrate con i software gestionali SAP, che permettono di svolgere determinate funzioni nel campo della Business Intelligence. Jobrapido utilizza il Data Warehouse Vertica di HP, e tramite SAP B.O. è possibile estrarre i dati utili all'analisi. Le varie operazioni di ETL e di analisi di mining sono state effettuate mediante i software open source **Knime e Weka**, utilizzando talvolta Excel per alcune analisi statistiche.

Knime è una piattaforma per l'analisi dei dati e per il reporting open source, che integra vari componenti per l'apprendimento automatico e il Data Mining. Un'interfaccia grafica permette all'utente il montaggio di nodi per la pre-elaborazione dei dati (ETL: Extraction , Transformation , Loading) , per la modellazione, l'analisi dei dati e la visualizzazione. Weka è un software per l'apprendimento automatico open source, sviluppato nell'Università di Waikato, in Nuova Zelanda, molto simile per funzionalità a Knime, ma con un interfaccia grafica diversa e meno incentrato sul processo di ETL.
Le visualizzazioni dei dati, dei risultati, e la creazione di reports sono stati implementati

utilizzando il software **Tableau**, sviluppato da un'azienda Americana relativamente alla data visualization nel campo della Business Intelligence (non si tratta di un software Open Source).

4.2 Struttura dell'analisi

Il processo che porta alla creazione di nuovi profili è articolata in tre parti distinte:

- **nella prima fase** si è effettuata un'analisi di clustering basata sulla **frequenza di accesso dell'utente**, costruendo quindi delle metriche ad hoc che permettano di calcolarne la sua frequenza. La difficoltà principale era quella di ottenere dei valori che riuscissero in pieno a descrivere il comportamento dell'utente, ogni qualvolta visitava il sito di Jobrapido, andando a misurarne la sua frequenza di accesso;

- **nella seconda fase** si è effettuata un'analisi di clustering basata sull'utilizzo per accesso da parte dell'utente. Anche in questo caso la difficoltà maggiore è stata riscontrata nel capire quali valori e attributi potessero descrivere e misurare al meglio, attraverso la costruzione di metriche appropriate, il grado di utilizzo del servizio da parte dell'utente, se, ad esempio, si tratta di un semplice "sguardo veloce" al portale oppure se l'utente si è soffermato e ha utilizzato alcune funzioni;

- **nella fase finale** si è effettuata l'ultima analisi di clustering, in cui si sono uniti entrambi gli approcci, clusterizzando perciò su tutte le metriche costruite fino ad ora. Questa fase finale, come è già stato descritto in precedenza, è stata eseguita per poter ottenere una profilazione completa del jobseeker e quindi arrivare al risultato finale di riuscire ad etichettare ogni singolo utente, basandosi esclusivamente sulla sua frequenza e il suo utilizzo del portale.

Nei paragrafi successivi vengono descritte le tre fasi, partendo dall'elaborazione dei dati e la costruzione delle metriche, fino alla scelta dell'algoritmo di clustering e all'interpretazione dei risultati.

4.3 Prima fase: la frequenza di accesso

La prima fase dell'analisi consiste nella segmentazione degli utenti basandosi esclusivamente sulla loro frequenza di accesso nei tre mesi presi in esame. Il primo passo è stato quello di costruire delle metriche adatte a tale scopo. E' importante sottolineare il fatto che, per poter uniformare i valori, l'intervallo di tempo osservato per ogni utente è "personalizzato", nel senso che le metriche che richiedono il calcolo del tempo in cui un utente è stato attivo utilizzano un intervallo temporale pari al reale periodo di attività dell'utente, che non è esattamente 3 mesi, ma dipende dalla data di conferma. Questa operazione è stata necessaria per evitare di "svantaggiare" gli utenti iscritti dopo il 1 gennaio. In questo modo un utente registratosi il 31/01 avrà un orizzonte di osservazione

di due mesi e non di tre, e le sue azioni saranno rapportate ad un periodo di attività reale. Dopo una serie di studi e analisi, si è deciso di utilizzare le seguenti tre metriche:

$$\textbf{ActivityRate} = \frac{(Num\_Distinct\_VisitDay)}{(Interval(Fixed))}$$

Questa metrica è stata calcolata rapportando il numero distinto di giorni di visita di ogni jobseeker per il suo personale orizzonte temporale.E' un valore percentuale che indica quanto tempo un utente è stato attivo rispetto all'intervallo di tempo osservato.
Quindi se un utente ha un AR pari al 100%, vuol dire che per tutto il periodo di tempo considerato, l'utente è entrato nel sito di Jobrapido ogni giorno. Ciò indicherà sicuramente una frequenza di accesso altissima.

$$\textbf{AVG Days beetween DayVisit} = \frac{(Sum(Interval\_btw\_two\_DistinctDays))}{(Num\_Distinct\_VisitDay - 1)}$$

Questo valore è stato calcolato sommando tutti i giorni che intercorrevano tra una visita e la successiva per ogni jobseeker, rapportandolo al numero distinto di giorni di visita meno uno (per correttezza l'ultimo giorno di visita non deve essere calcolato, in quanto sommiamo gli intervalli tra un giorno e l'altro).
Questa metrica esprime una media dei giorni che intercorrono tra una visita e l'altra. Più il valore tende verso 1, più il jobseeker è tornato di frequente, quasi ogni giorno.

$$\textbf{Visit Per Day} = Avg\_Visits\_per\_Day$$

E' un valore che esprime la media del numero di visite giornaliere di ogni utente, in quanto in uno stesso giorno l'utente può compiere più visite. Nello specifico, il sistema conteggia le visite nel seguente modo:

- dopo 30 minuti di inattività da parte dell'utente, la visita termina;
- se l'utente naviga la notte, passata la mezzanotte, scatterà in automatico il conteggio di una nuova visita;
- se l'utente abbandona il sito, e poi rientra, la visita successiva sarà considerata nuova.

Queste tre metriche esprimono in modo puntuale la frequenza di un utente per accesso sul portale di Jobrapido.
Un Jobseeker caratterizzato, per esempio, da un valore di AR = 100%, da un AVG=1 e da un VPD = 5, sarà un utente attivissimo, in quanto nell'intervallo di tempo osservato è entrato tutti i giorni con una media di 5 visite al giorno.

Nel paragrafo seguente viene illustrata l'attività di **data preparation** e **data quality** che hanno permesso di costruire le metriche descritte in precedenza.

4.3.1 Data preparation e data quality prima fase

La fase di **Data Preparation** mira alla preparazione dei dati al fine di essere utilizzati correttamente per la creazione delle metriche, mentre la fase di **Data Quality** ha lo scopo di migliorare la qualità dei dati per poter ottenere delle analisi più precise e accurate. Sono stati analizzati nel dettaglio vari aspetti critici che spesso i dati presentano al loro interno e sono stati gestiti secondo opportune scelte, ritenute rilevanti al fine delle analisi.
I vari step che hanno seguito la data preparation e la data quality sono i seguenti:

- **missing values:** La presenza di missing values all'interno di un dataset è un evento comune, ed è necessario trattarli utilizzando la strategia più appropriata. Le cause possono essere molteplici: i dati non erano disponibili, i dati non erano stati registrati perchè irrilevanti, i dati erano stati dimenticati o erroneamente cancellati; oppure l'evento non era mai accaduto.
  Esistono diverse strategie di trattamento dei missing value come: ignorare i valori mancanti, escludere tutti i record che li contengono, sostituirli con altri valori come la media o la mediana, dedurre i valori mancanti utilizzando valori esistenti.
  In generale, bisogna scegliere la strategia più appropriata nel contesto in cui si opera.
  In questo caso, all'interno dei dataset presi in esame, erano presenti dei valori mancanti negli attributi riferiti al campo *Email* e al campo *Date/Time* della visita effettuata dall'utente. Essendo un numero molto ridotto rispetto al campione esaminato (circa il 2% del dataset completo), sono stati eliminati accuratamente tutti quei dati riferiti a utenti a cui non è stato possibile associare una mail e/o i loro giorni di visita.
  In questo modo si è lavorato solo con utenti che era possibile riconoscere ed identificare da un indirizzo mail valido, in quanto chiave univoca;

- **outliers:** possono essere definiti come valori di un attributo che sono inusuali, oppure molto differenti rispetto a quelli assunti tipicamente. L'individuazione dell'eventuale presenza di outliers è avvenuta studiando la distribuzione dei singoli attributi mediante istogrammi e box plot.
  La **box plot** è un metodo che mostra la distribuzione dei valori di un singolo attributo numerico basandosi sui percentili, evidenziando la presenza di outliers. In questo contesto sono stati individuati numerosi outliers in diversi attributi riferiti al numero di visite totali e giornaliere.
  Spesso questi valori erano molto elevati rispetto ai valori comunemente analizzati. L'idea principale è quella relativa alla presenza di crowler, che visitando il sito Jobrapido provenendo da terze parti, effettuavano un numero di visite giornaliere talmente elevato che risultava difficile pensare all'azione di un utente umano. Per cercare di ovviare a questo problema, si è deciso di approcciarsi in modo scientifico, utilizzando la tecnica statistica dello **Z-score**.

La standardizzazione è un procedimento che riconduce una variabile aleatoria distribuita secondo una media μ e varianza σ^2, a una variabile aleatoria con distribuzione "standard", ossia di media zero e varianza pari a 1.
I valori standardizzati, mediante lo Z-score, possono essere utilizzati per identificare gli outlier, in quanto si può affermare che, per i dati che hanno una distribuzione a campana, quasi tutti i valori si troveranno entro tre-cinque deviazioni standard dalla media.
Per questo motivo si è deciso di trattare come outlier tutte le osservazioni con uno z-score inferiore a -5 o maggiore di +5. Il calcolo dello Z-Score per le visite giornaliere è stato il seguente:

- ottenere la media del numero di visite per giorno di tutti i jobseeker analizzati;

$$y = \frac{(Sum(Visit\ Day\ di\ tutti\ i\ JS))}{(totalRows)}$$

- determinare σ, mettendo in una sommatoria tutti i valori delle visite per giorno di ogni utente, dove X_i corrisponde al numero di visite giornaliere di ogni utente:

$$\sigma = \frac{\sqrt{\sum (X_i - y)^2}}{total\ rows}$$

- determinare per ogni X_i se il suo valore sia $> +5$ o < -5, in modo tale che tali valori vengano filtrati ed eliminati.

$$\frac{|(X_i - y)|}{\sigma} > 5$$

  In questo modo è stato possibile individuare tutti quei valori ritenuti anomali, che avrebbero potuto influire negativamente sulla bontà dei risultati;

- **duplicate rows and inconsistence:** un aspetto molto importante mira all'individuazione di valori duplicati o eventualmente inconsistenti, e quindi con valori decisamente errati rispetto a quelli che ogni attributo dovrebbe normalmente assumere.
  In questo contesto non sono stati riscontrati record duplicati, mentre sono stati riscontrati alcuni attributi con valori decisamente errati come alcune date relative ai giorni di visita (per alcuni errori di inserimento erano presenti delle date di visite nella forma 9/9/1999). In questi casi non si è tenuto conto del dato.

Tutto ciò ha determinato una lunga fase iniziale, necessaria per poter preparare i dati in modo adeguato alle future analisi e per poter migliorare la bontà dei risultati.

Una volta che i dati sono stati resi pronti per le analisi, si è iniziato a lavorare sul tipo di **algoritmo** da utilizzare. Come già illustrato nei capitoli precedenti, la scelta è ricaduta su due, tra i più diffusi nelle analisi di clustering: **DBSCAN e K-Means**. L'idea di

partenza è stata quella di eseguire le analisi con entrambi gli algoritmi, confrontare i risultati ottenuti e scegliere quello migliore.

4.3.2 K-Means prima fase

L'algoritmo K-means è una tecnica di clustering partizionale che permette di suddividere un insieme di oggetti in K gruppi sulla base dei loro attributi. In sintesi, il procedimento è il seguente: inizialmente viene deciso a priori un numero K di cluster. Ogni cluster viene associato a un centroide (la media aritmetica di un gruppo di punti) e in modo iterativo ogni punto viene assegnato al relativo centroide più vicino, in modo tale che ogni insieme di punti associati allo stesso centroide formi un cluster. A ogni passo vengono ricalcolati i centroidi, fino a quando l'algoritmo non converge. Essendo un approccio partizionale, ogni punto potrà appartenere ad un solo cluster.

Lo scopo è quello di raggruppare tutti gli utenti selezionati per determinare diversi profili basati sulla frequenza di accesso al sito.

La soluzione trovata dall'algoritmo k-means dipende dalla scelta del K, e dunque il primo passo è stato quello di determinare il numero più appropriato.

Per valutare la bontà della scelta del valore K, la letteratura illustra come sia possibile utilizzare il valore del **parametro SSE**, definito come la **somma degli errori quadrati**, cioè la somma delle distanze al quadrato tra tutti i punti del cluster e il relativo centroide, calcolando successivamente la somma totale, secondo la formula:

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x)$$

dove x è un punto appartenente al cluster C_i, mentre m_i è il centroide del cluster C_i.

Per effettuare questa operazione è stato utilizzato il software **Knime**, tramite il nodo SimpleKMeans, che sviluppa l'algoritmo presente nel software Weka.

Sono stati eseguiti numerosi tentativi, cambiando ripetutamente sia il seed che il valore K in quanto l'algoritmo come punto di partenza sceglie in modo casuale un punto (in base al seed impostato) e, partendo da quello, itera, creando ogni volta delle nuove partizioni, fino a quando non converge.
Per questo motivo l'algoritmo risulta molto sensibile al seed, e valori differenti possono portare risultati diversi, talvolta anche con uno scostamento importante. Ciò ha permesso di contrastare il problema della scelta dei centroidi, in quanto vengono generati casualmente dall'algoritmo, in base al valore scelto del parametro seed. Dunque, per effettuare una buona analisi sono stati fatti numerosi tentativi, partendo da K=2 fino a K=10, variando il seed con più di 30 valori diversi.

Per ogni K è stata ricavata la media tra i diversi SSE ottenuti, utilizzati per la creazione di un grafico che mostra l'andamento dei valori.

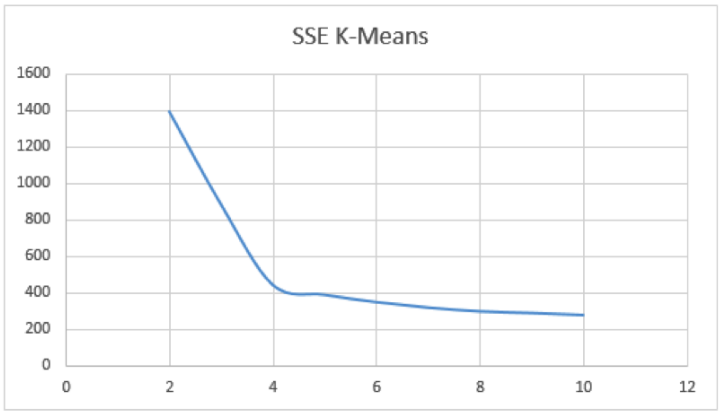

Figura 4.1: Grafico studio SSE

Il grafico mostra una funzione dove l'eventuale flesso potrebbe corrispondere al valore K più appropriato per effettuare il clustering.

Ovviamente, con l'aumentare del numero di cluster, l'SSE tende a diminuire. Ciò è dovuto alla tendenza dei clusters di ridurre la propria dimensione con l'aumentare del numero dei raggruppamenti, abbassando la somma degli errori quadrati; questo non permette necessariamente un miglioramento del clustering. Infatti, la condizione migliore è quella di riuscire a trovare un K che sia il più piccolo possibile, ma che allo stesso tempo presenti un basso SSE.

Perciò, dove si verifica un flesso, e quindi una diminuzione della pendenza, potrebbe risiedere la soluzione migliore. Dal grafico sopra riportato si può notare come l'andamento della curva risulti essere piuttosto regolare, presentando però un piccolo flesso.
Dopo un'attenta analisi, in conclusione, è stato scelto un valore di K uguale a 4 con seed pari a 32, punto nel quale appare la variazione di pendenza più evidente.

4.3.3 DBSCAN prima fase

Il **DBSCAN** è un metodo di clustering basato sulla densità, che permette di individuare regioni di punti con una densità sufficientemente alta, risultando efficace in dataset caratterizzati da outliers e noise.
Per effettuare l'analisi è stato utilizzato il software Knime, sfruttando i nodi DBSCAN e Weka Cluster Assigner del pacchetto Weka 3.7. Per la scelta dei parametri ottimali con cui eseguire l'algoritmo si è condotto uno studio attraverso un grafico delle distanze tramite Open Office.

Nel DBSCAN i **parametri** su cui poter variare l'esecuzione sono due:

- **eps:** il raggio entro cui valutare la densità di un punto;
- **minPoint:** il minimo numero di punti nell'intorno di quello considerato, che permette di classificarlo come core-point, vale a dire come facente parte di una regione sufficientemente densa, o come punto di noise.

L'approccio per determinare i due parametri in questione è basato sull'osservare il comportamento della distanza (k-dist) di un punto dal suo k-esimo più vicino. Il primo passo è quello di stabilire il parametro minPoint: per analisi bidimensionali è ragionevole scegliere come parametro il valore 4, ma è possibile cercare empiricamente un valore più soddisfacente, a seconda dei casi, della densità e della distribuzione degli oggetti. In ogni caso, il valore opportuno di eps viene scelto come conseguenza del minPoint.

Per stabilire il **corretto valore** di eps si calcola la distanza di ogni punto dal suo k-esimo punto più vicino (k-dist nearest neighbour). Successivamente, si costruisce il grafo delle distanze mettendo sull'asse delle ascisse i punti ordinati con distanze crescenti e sull'asse delle ordinate i valori delle distanze.

Per costruire la **matrice delle distanze**, per ragioni computazionali, sono state utilizzate alcune funzioni del linguaggio di programmazione R. La matrice esportata è stata poi elaborata su Calc, che ha permesso di calcolare per ogni punto il k-esimo elemento più vicino attraverso la funzione PICCOLO() e la relativa k-dist. Con un riordinamento degli oggetti per valori crescenti di k-dist, si sono potuti realizzare i grafici delle distanze.

E' importante sottolineare che è stato necessario preliminarmente normalizzare gli attributi per poter poi realizzare i grafici, in quanto il nodo DBSCAN si aspetta un eps che agisca su valori normalizzati.
In questo caso, però, i risultati non sono stati soddisfacenti, in quanto l'algoritmo, pur variando i parametri, non riusciva a ottenere una buona partizione. Spesso non riusciva a determinare più di un cluster, e questo era dovuto molto probabilmente alla eccessiva concentrazioni dei punti nello spazio, creando una sola zona densa, difficile da suddividere per un algoritmo basato sulla densità.
Alla fine è stato preferito l'algoritmo K-Means per le ragioni illustrate precedentemente. I risultati generati vengono riportati e descritti di seguito.

4.3.4 Risultati prima fase

Di seguito vengono riportate una serie di visualizzazioni ottenute con il software Tableau, i risultati ottenuti, in dettaglio con le loro interpretazioni.

Dopo la scelta dell'algoritmo, e la sua applicazione con i parametri studiati, si sono

ottenuti **4 profili distinti** di utenti sulla base della loro frequenza di accesso al sito. Il primo grafico a torta riporta la distribuzione dei cluster a seguito dell'analisi, con le rispettive percentuali riferite al dataset totale. Dall'analisi sono emersi 4 gruppi distinti, e sono stati denominati F1, F2, F3, F4, dove il primo è il gruppo che raccoglie gli utenti con una frequenza più alta, via via abbassandosi fino al gruppo F4.

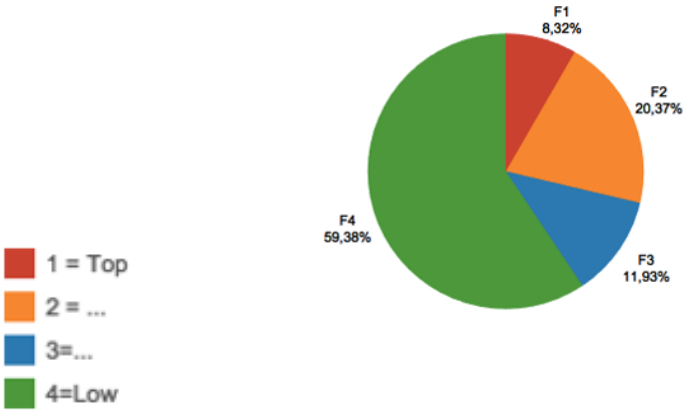

Figura 4.2: Diagramma a torta profili primo cluster

Il dataset analizzato conteneva esattamente 92.634 utenti distinti, dopo le opportune operazioni di data preparation e di data quality.
La tabella successiva, invece, evidenzia per ogni metrica e per ogni cluster il valore corrispondente al centroide (che in questo caso corrisponde al valore medio) che l'algoritmo di cluster ha trovato.

Clusters	Avg. AR	Avg. AVG	Avg. Visit Per Day
F1	0.646	1.573	1.635
F2	0.302	2.859	1.510
F3	0.107	5.618	2.701
F4	0.087	5.975	1.471

Figura 4.3: Tabella con i valori medi per metrica del primo cluster

Come si può notare, il gruppo F1 è caratterizzato da utenti che in media hanno un

activity rate molto alto (circa il 64%), e mediamente ogni utente appartenente a questo cluster torna circa ogni giorno e mezzo.
Il gruppo F4, invece, è caratterizzato da un activity rate medio molto basso (pari al 0,08%), e questo vuol dire che gli utenti, spesso, hanno effettuato solo una o due visite in totale.
La terza rappresentazione riporta una tabella con i valori di massimo e di minimo di ogni metrica per ogni cluster determinato.

Clusters	Min. AR	Max. AR	Min. AVG	Max. AVG	Min. Visit Per Day	Max. Visit Per Day
F1	0.449	1.000	1.0	3.00	1.00	4.714
F2	0.182	0.481	1.0	6.89	1.00	2.682
F3	0.027	0.479	1.0	69.00	2.08	4.750
F4	0.027	0.204	1.0	70.00	1.00	2.000

Figura 4.4: Tabella con il Range Max-Min delle metriche primo cluster

Da quest'ultima tabella si può evincere, invece, come la metrica relativa all'activity rate sia molto ben discriminante tra i vari cluster, suddividendosi in modo molto preciso. I restanti range hanno delle sovrapposizioni, ma ciò non significa che i risultati siano negativi. In questo caso ogni cluster è caratterizzato da un preciso range e da un determinato centroide, e ciò permette di definire in modo concreto diversi profili di utenti basandosi sulla loro frequenza. Infatti dai tre grafici sopra riportati si può notare come siano emersi quattro profili diversi, basati sulla frequenza di accesso al sito.

Questa è una breve interpretazione dei 4 profili di utenti determinati dalla prima fase dell'analisi di clustering:

- **gruppo F1** (9% del dataset)
  Appartengono a questo profilo tutti gli utenti con un'alta frequenza di accesso al sito. Possono essere considerati gli utenti migliori. Sono caratterizzati da una AVG Days btw dayvisit tendente ad 1, e nell'intervallo analizzato, sono tornati molto spesso (presentano un AR ¿ 50%);

- **gruppo F2** (20% del dataset)
  Questo cluster identifica gli utenti con una frequenza di accesso medio/alta, ma con una media di giorni che intercorrono tra una visita e l'altra minore del gruppo F3;

- **gruppo F3** (12% del dataset)
  In questo gruppo, gli utenti sono caratterizzati da una frequenza di accesso al sito

web medio/bassa, ma in media la metrica Visit per Day risulta essere la più alta in assoluto;

- **gruppo F4** (59% del dataset)
  Questo cluster identifica gli utenti che sono caratterizzati da una frequenza di accesso molto bassa. Dopo la prima visita, gli utenti di questo gruppo ritornano in media sul portale dopo circa un mese o più.

Di seguito riportiamo per una migliore comprensione alcune distribuzioni relativamente alle metriche utilizzate per l'analisi di clustering.

La prima distribuzione pone sull'asse x la metrica AVG Days Btw Dayvisit, mentre sull'asse delle y AR.
In questo caso (filtrando il cluster F3) si può notare come i gruppi di utenti vengano ben separati, in quanto l'activity rate risulta essere un attributo ben discriminatorio nella creazione dei profili.

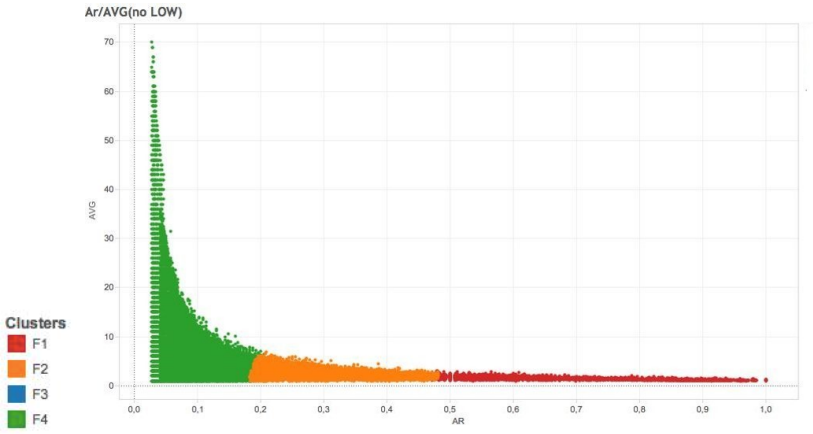

Figura 4.5: Grafico che illustra lo Scatter Plot tra Activity Rate e AVG per cluster (filtrato cluster F3)

La sovrapposizione è più evidente nel momento in cui si considerano tutti e 4 i gruppi distinti. Resta comunque una buona separazione tra clusters.

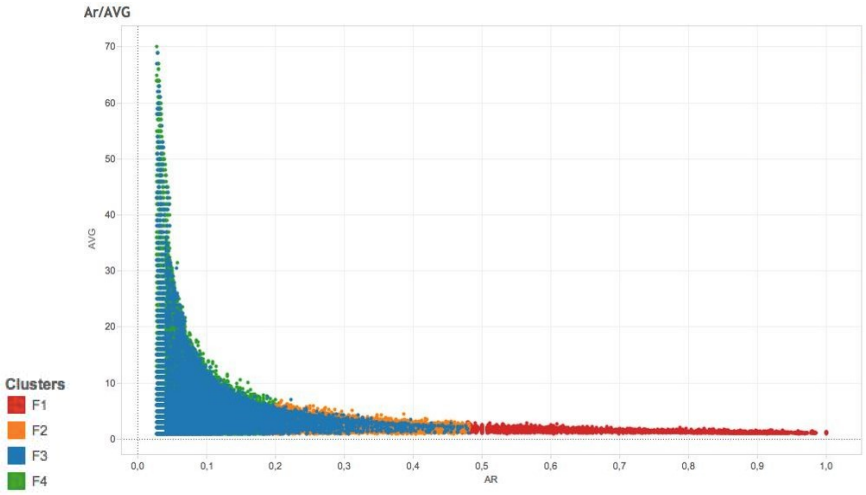

Figura 4.6: Grafico che illustra lo Scatter Plot tra Activity Rate e AVG per cluster

In questa rappresentazione sono stati messi a confronto AR e VisitPerDay. Anche in questo caso la separazione effettuata dall'algoritmo K-Means ha avuto dei buoni risultati, formando cluster ben separati.

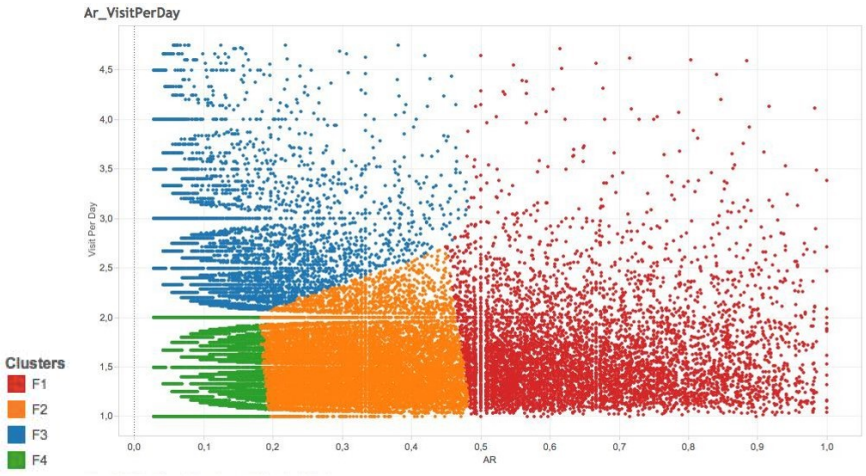

Figura 4.7: Scatter Plot tra Activity Rate e Visit Per Day per cluster

Infine, si è deciso di riportare, sempre per una migliore comprensione dell'elaborato, una rappresentazioni 3D, ottenuta utilizzando il linguaggio di programmazione R.

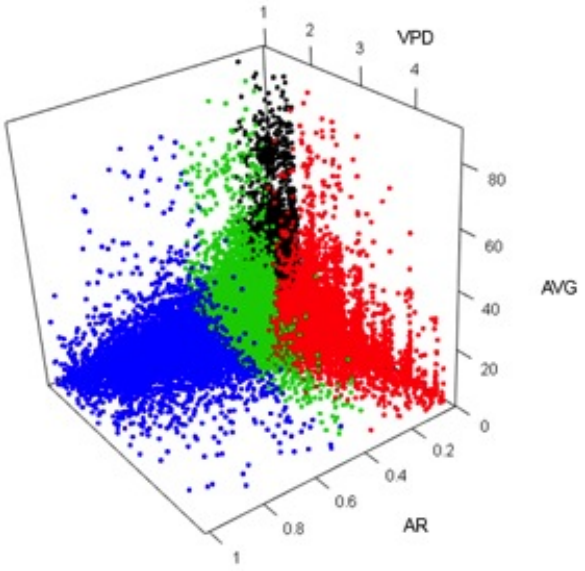

Figura 4.8: Rappresentazione 3D della prima analisi di cluster

La figura illustra in modo 3D come i punti si posizionino nello spazio, rispetto alle tre metriche utilizzate per l'analisi di clustering. Questo permette di mostrare in modo più evidente il buon risultato ottenuto con la prima analisi. Il colore blu rappresenta gli utenti F1, in quanto sono collocati nella parte superiore dell'asse AR perché caratterizzati da un activity rate molto alto. Il colore verde rappresenta gli utenti F2, il colore rosso gli utenti F3, mentre il colore nero gli utenti F4, in quanto sono caratterizzati da alti valori di AVG Days beetween DayVisit.

4.4 Seconda fase: il tasso di utilizzo

Nella seconda fase ci si è invece focalizzati sulla profilazione degli utenti, basandosi sul **tasso di utilizzo** del sito web, inteso come numero di pagine visitate durante la navigazione, durata di ogni visita e tasso di abbandono.
In questo caso si è cercato di costruire delle metriche che permettessero di descrivere la "quantità di utilizzo" che ogni utente compie durante le sue visite. In pratica si è cercato di misurare quanto un utente usa il portale a ogni visita, se sfrutta tutte le potenzialità e i servizi offerti da Jobrapido, oppure effettua visite molto veloci, dando un semplice "sguardo".

Dopo una serie di studi e analisi, si è deciso di utilizzare le seguenti tre metriche:

$$\textbf{AVG Num Page Per Visit} = \frac{(Num\_Total\_Page\_Views)}{(Num\_Total\_Visits)}$$

Rappresenta una media delle pagine visitate per visita. Perciò questa metrica indica, in media, quante pagine ogni singolo utente ha visto, rispetto al totale delle visite effettuate

su Jobrapido.
E' un valore che può iniziare a esprimere il tasso di utilizzo del sito internet, ma da solo non basta.
Infatti, un utente con un'alta media di pagine viste, probabilmente sarà un tipo di utente caratterizzato da un tasso di utilizzo elevato, ma non è sempre detto, in quanto queste pagine potrebbero essere state viste in poco tempo, senza usufruire a pieno dei servizi offerti dal portale.

$$\textbf{AVG Time on site Per Visit} = \frac{(Sum(Time\_on\_visit))}{(Num\_Total\_Visits)}$$

Questo valore esprime il tempo trascorso, in secondi, sul sito di Jobrapido rispetto al totale delle visite effettuate. Un utente caratterizzato da un tasso di utilizzo elevato, potrà avere una media relativa al tempo molto alta, in quanto avrà effettuato delle visite soffermandosi per più tempo sulle varie pagine, e quindi potrebbe essere profilato come un utente molto attivo dal punto di vista dell'utilizzo.

$$\textbf{Bounce Rate} = \frac{(Num\_Visits\_with\_0\_clicks\_and\_1\_page\_view)}{(Num\_Total\_Visits)}$$

Questa metrica non rappresenta il classico bounce rate utilizzato solitamente per effettuare questo tipo di analisi, ma è un valore che è stato riadattato alle esigenze della nostra analisi.
In questo caso per bounce rate si intende la percentuale di visite in cui l'utente non ha fatto alcun tipo di click e ha visitato una sola pagina, rapportato al numero totale di visite.
Ciò significa che tutti quegli utenti che hanno visualizzato in una visita una sola pagina e non hanno effettuato alcun tipo di interazione, molto probabilmente non hanno trovato interessanti i risultati ottenuti, e quindi potrebbe rivelarsi un buon indice descrittivo.

Queste metriche sono in grado di esprimere il grado di utilizzo del portale da parte degli utenti. Per esempio, un utente potrà essere caratterizzato da un tempo medio di visita alto e da un basso bounce rate, e ciò significa che l'utente utilizza molto i servizi offerti, cliccando e visitando più pagine.

4.4.1 Data preparation e data quality seconda fase

Come precedentemente descritto nel paragrafo relativo alla prima fase di analisi, anche in questo caso è stata necessaria la fase di **Data preparation** e di Data quality per poter preparare e pulire i dati con lo scopo di creare le metriche illustrate nel paragrafo precedente.
Poiché si sono utilizzati gli stessi dati e gli stessi attributi, anche se sono stati aggregati in modo diverso per formare le nuove metriche, gli errori, gli outlier e i missing values riscontrati sono stati gli stessi della prima fase. Per quanto riguarda gli outlier, anche questi sono stati identificati e successivamente filtrati mediante il procedimento dello

z-score, riscontrandoli principalmente nell'attributo relativo al numero di visite.
Per quanto riguarda l'attributo legato al tempo di visita, questo presentava spesso alcune inconsistenze o dei valori mancanti. Ma essendo una percentuale molto bassa rispetto al totale (circa il 3% del dataset finale) sono stati opportunamente eliminati.

4.4.2 K-means seconda fase

Anche in questa seconda fase, a seguito degli studi effettuati nella prima fase, si è deciso di usare **l'algoritmo k-means** per poter segmentare gli utenti sulla base del loro tasso di utilizzo.
Di seguito viene riportato lo studio eseguito per poter scegliere il K ottimale, da impostare successivamente per l'esecuzione dell'algoritmo.

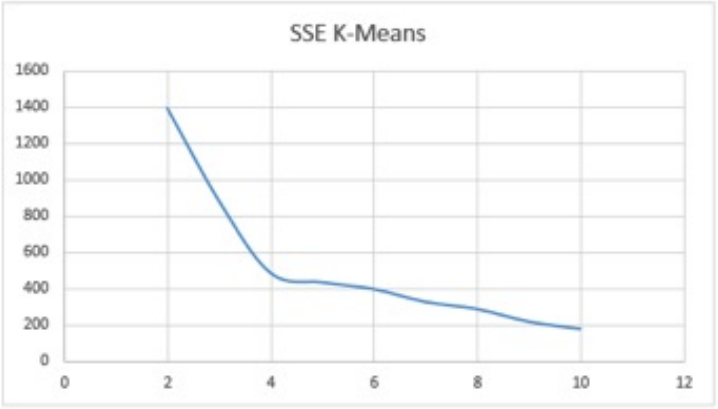

Figura 4.9: Grafico studio SSE

Anche in questo caso, per effettuare una buona analisi, sono stati fatti numerosi tentativi, partendo da K=2 fino a K=10, variando il seed con più di 30 valori diversi. Il risultato è mostrato nel grafico precedente, in cui è possibile notare il flesso ricercato sempre nel punto con K = 4. In definitiva si sono quindi trovati 4 profili diversi basati sul tasso di utilizzo del portale. Nel paragrafo successivo verranno mostrati graficamente i risultati ottenuti con le relative interpretazioni.

4.4.3 Risultati seconda fase

Dopo aver effettuato l'analisi di clustering, sono emersi **quattro profili** diversi, denominati U1, U2, U3, U4, dove il primo è il gruppo che raccoglie gli utenti con un tasso di utilizzo elevato, via via diminuendo fino al gruppo U4. Di seguito viene riportato il grafico a torta con il risultato della profilazione e le rispettive percentuali sul dataset totale.

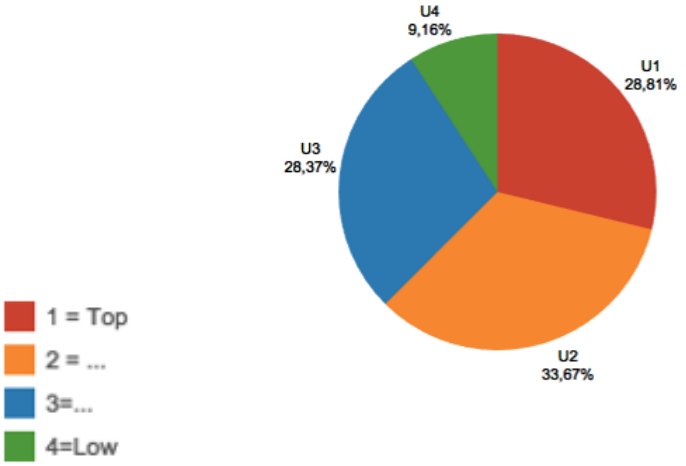

Figura 4.10: Diagramma a torta dei profili secondo cluster

A differenza del risultato precedente, in questo caso il cluster U4, relativo agli utenti con il minor tasso di utilizzo, ha una percentuale di popolazione nettamente inferiore all'analisi precedente sulla frequenza di accesso, e soprattutto è il minore dei 4 gruppi trovati. Viceversa, il gruppo U1, che raccoglie gli utenti con un alto tasso di utilizzo, e quindi molto preziosi per l'azienda, occupa circa il 30% del dataset totale.

Di seguito vengono riportati sia i valori relativi ai centroidi, sia i range di massimo e di minimo per ogni metrica e per ogni cluster, in modo tale da illustrare, in modo più preciso e accurato, la distribuzione dei valori all'interno dei differenti profili.

Clusters	Avg. AVG NumPagesPerVisit	Avg. AVG Time On Site Per Visit	Avg. Bounce Rate
U1	3.22	153.76	0.03
U2	2.73	129.44	0.21
U3	2.24	103.08	0.41
U4	1.70	60.27	0.65

Figura 4.11: Tabella con i valori medi per metrica del primo cluster

Dalla prima tabella si può notare come la media del Bounce Rate relativa al cluster U1 sia molto bassa rispetto ai restanti valori presenti negli altri profili. Mentre, il cluster U4 è ben distinto dagli altri gruppi per i suoi valori medi molto distanti ed elevati.
Da questa tabella è possibile dedurre come gli utenti presenti nel cluster U1, in media, guardano circa 3 pagine e mezzo a ogni visita, spendono circa quasi 3 minuti a visita e hanno un tasso di abbandono estremamente basso. Se ci spostiamo all'altra estremità, il gruppo U4 è caratterizzato da utenti che guardano quasi due pagine per ogni visita, ma il loro bounce rate è estremamente alto, circa il 65%. Questo significa che su 10 visite effettuate, 6,5 visite sono state caratterizzate dalla visione della sola prima pagina senza effettuare alcun tipo di interazione.
Questo può essere dovuto magari al fatto che non hanno trovato il contenuto che cercavano, o magari le parole chiave utilizzate non sono state quelle corrette. A tutte queste domande si cercherà di rispondere attraverso l'analisi di classificazione effettuata nella seconda parte.

Nella seconda tabella, invece, è importante sottolineare come il Bounce rate sia quasi ben distinto per ogni cluster, con minime sovrapposizioni di valori. Questo è anche visibile nelle visualizzazioni successive. Ciò ci permette di dedurre come la metrica del Bounce Rate sia stata molto discriminante nel creare i quattro profili esaminati.

Clusters	Min. AVGPages	Max. AVGPages	Min.AVGTime	Max.AVGTime	Min.BRate	Max.BRate
U1	1.056	27	0.389	2,510.0	0.00	0.16
U2	1.048	17	0.278	1,692.5	0.12	0.31
U3	1.053	20	0.074	2,945.8	0.31	0.53
U4	1.017	10	0.134	837.8	0.53	0.98

Figura 4.12: Tabella del Range Max-Min metriche secondo cluster

Di seguito si riporta una breve interpretazione dei cluster individuati:

- **gruppo U1** (29% del dataset)
  Appartengono a questo profilo tutti gli utenti con un alto tasso di utilizzo per accesso. La durata media per ogni visita è la più alta in assoluto tra i vari profili e il Bounce rate è il più basso. Mediamente spendono 3 minuti per ogni visita;

- **gruppo U2** (34% del dataset)
  Questo cluster identifica gli utenti con un tasso medio di utilizzo per accesso. A differenza del cluster U1, hanno un valore medio di Bounce Rate 7 volte più alto del gruppo U1;

- **gruppo U3** (28% del dataset)
  In questo gruppo, gli utenti sono caratterizzati da un Bounce Rate doppio rispetto

al Bounce rate medio del gruppo U2. Inoltre, la durata media di ogni visita è minore di circa 30 secondi rispetto al gruppo U2. In generale, hanno un tasso medio/basso di utilizzo per accesso;

- **gruppo U4** (9% del dataset)
  Questo cluster identifica gli utenti che sono caratterizzati da un tasso di utilizzo estremamente basso. La durata media di visita è circa di 1 minuto, mentre le pagine visitate in media per accesso sono circa 1-2. Mediamente, il tasso di abbandono è pari al 65% del periodo osservato.

Queste due fasi distinte hanno permesso di segmentare gli utenti che hanno utilizzato Jobrapido Italia, basandosi sulla loro frequenza di accesso e sul loro tasso di utilizzo. In questo modo è stato possibile ottenere quattro profili di utenti, che permettono di poterli descrivere in modo preciso e puntuale sulla base di specifiche metriche, assegnandoli una precisa etichetta, necessaria per la seconda grande analisi legata alla classificazione. Il passaggio finale è stato quello di unire le due analisi in una fase finale, utilizzando tutte e 6 le metriche costruite. Le motivazioni che hanno portato alla creazione di un cluster finale sono le seguenti:

- fornire una prima descrizione di quali fossero le tipologie di utenti che utilizzano il sito di Jobrapido per poi arrivare a fornire un'etichetta unica ad ogni JS, in modo tale da utilizzare un profilo definitivo nelle fasi di analisi successive;
- individuare quali combinazioni possono emergere, utilizzando le metriche legate alla frequenza e al tasso di utilizzo. Questo è il cluster più importante per capire che genere di utenti utilizzano il portale e le possibili combinazioni.

Infine, anche in questa seconda fase, si riporta, ai fini di una migliore comprensione dell'elaborato, una rappresentazione 3D ottenuta utilizzando il linguaggio di programmazione R, che mostra il risultato della seconda analisi di clustering.

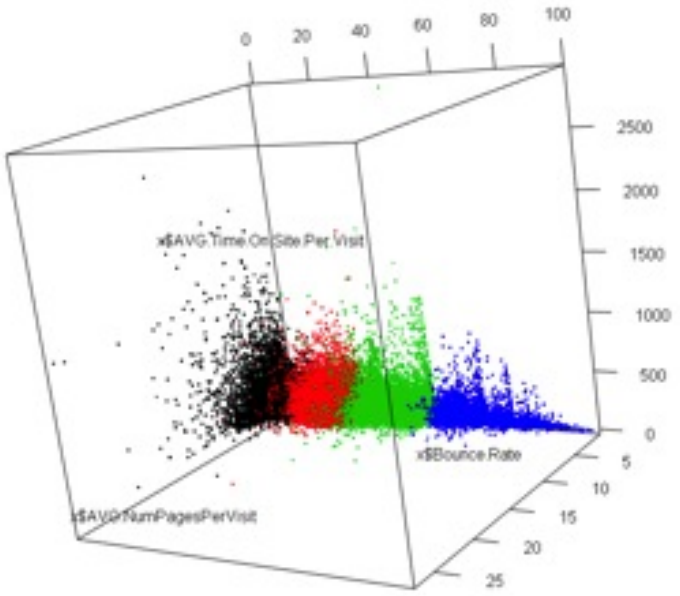

Figura 4.13: Rappresentazione 3D secondo cluster

Anche in questo caso è facile intuire come gli utenti caratterizzati dal colore nero siano i jobseeker appartenenti al gruppo U1, in quanto presentano valori molto bassi di bounce rate. Il colore rosso rappresenta gli utenti del gruppo U2, il colore verde quelli del gruppo U3 e infine il colore blu identifica gli utenti del gruppo U4. Una rappresentazione 3D riesce a migliorare la comprensione del risultato di clustering, e facilita la lettura della posizione degli utenti nello spazio, rispettivamente sui tre assi che rappresentano le tre metriche utilizzate.

4.5 Fase finale

La fase finale consiste in una definitiva e **ultima analisi di clustering** in cui si è deciso di unire tutte e sei le metriche utilizzate nelle due precedenti fasi. Questo è il punto più importante, in quanto ci permette di attribuire a ogni utente analizzato un'etichetta definitiva sulla base del suo comportamento. Questa etichetta sarà il punto di partenza nella fase di classificazione, in quanto si andranno a indagare le motivazioni comportamentali e/o di esperienza avute sul sito, che hanno portato gli utenti a fare parte di un gruppo piuttosto che di un altro.
Anche in questo caso, si è mantenuto l'utilizzo dell'algoritmo K-means, che si è rivelato l'algoritmo migliore per i suddetti scopi.

4.5.1 Risultati fase finale

Nella prima rappresentazione viene mostrato graficamente il risultato dell'analisi di clustering, che ha evidenziato quattro profili definitivi sulla base della frequenza di accesso e del tasso di utilizzo del portale.

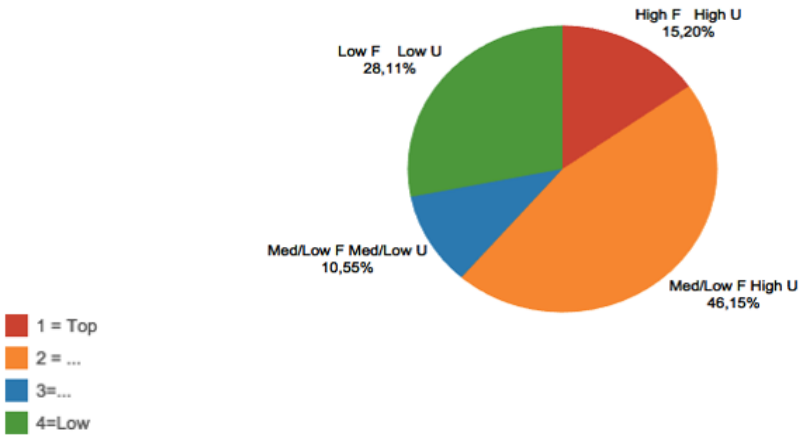

Figura 4.14: Diagramma a torta dei profili del cluster finale

Come si può notare, il profilo che raccoglie gli utenti migliori è quello rosso (High F High U) caratterizzato da utenti con un alto tasso di utilizzo e un'alta frequenza di accesso. Sono circa il 15% del totale del dataset e sicuramente può essere considerata una buona percentuale.
Il gruppo verde è l'estremo opposto: rappresenta utenti con valori pessimi sia per la frequenza che per l'utilizzo. I due gruppi intermedi, invece, si distinguono solo per il tasso di utilizzo (alto per il cluster arancione e basso per il cluster blu), mentre presentano valori molto simili per quanto riguarda la frequenza di accesso.

Per poter studiare in modo più approfondito i valori che caratterizzano ciascun profilo, si riportano di seguito alcune delle visualizzazioni effettuate per comprendere meglio i dati a disposizione.
Nella tabella seguente vengono riportati per ogni cluster e per ogni metrica i valori medi (centroidi) corrispondenti, individuati dall'algoritmo k-means.

Winner Cluster	Avg. AR	Avg. AVG	Avg. Visit Per Day	Avg. AVG NumPagesPerVisit	Avg. AVG Time On Site Per Visit	Avg. Bounce Rate
High F High U	0.541	1.869	1.585	2.990	130.428	0.201
Med/Low F High U	0.129	4.994	1.435	2.929	137.998	0.116
Med/Low F Med/Low U	0.100	5.820	2.584	2.553	130.707	0.300
Low F Low U	0.118	5.923	1.486	2.029	83.089	0.472

Figura 4.15: Tabella dei valori medi per metrica del cluster finale

Dalla tabella si può notare come il profilo **High F High U** sia caratterizzato in media da un alto **activity rate** e da un **bounce rate** relativamente basso: gli utenti, in media, spendono 2,5 minuti per visita. I due profili intermedi presentano più o meno valori molto vicini, fatta eccezione del bounce rate, più alto nel terzo gruppo, e la media delle visite per giorno, più alta sempre nel terzo gruppo. Infine il profilo **Low F Low U** presenta i valori peggiori, in quanto rappresenta gli utenti che non hanno né un'alta frequenza di accesso, né un alto tasso di utilizzo del sito.

Di seguito vengono riportati alcuni tra i grafici più significativi che mostrano le relative distribuzioni delle varie metriche rispetto ai profili individuati.

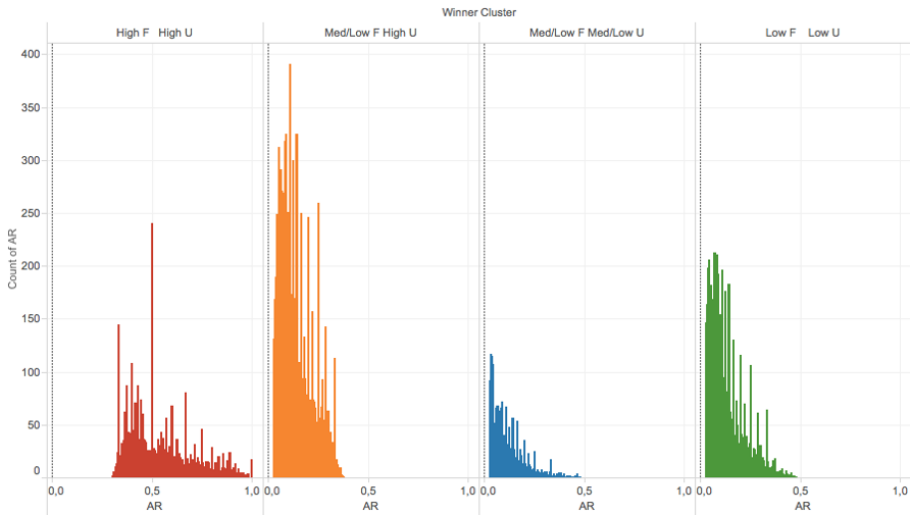

Figura 4.16: Istogramma relativo alla distribuzione della metrica activity rate

In questo istogramma che descrive la distribuzione della metrica activity rate, si nota

come il gruppo 1 (rosso) sia caratterizzato da utenti con valori molto alti (in quanto sono tutti gli utenti con una frequenza molto alta), mentre via via i valori dell'activity rate diminuiscono, cambiando i vari clusters.

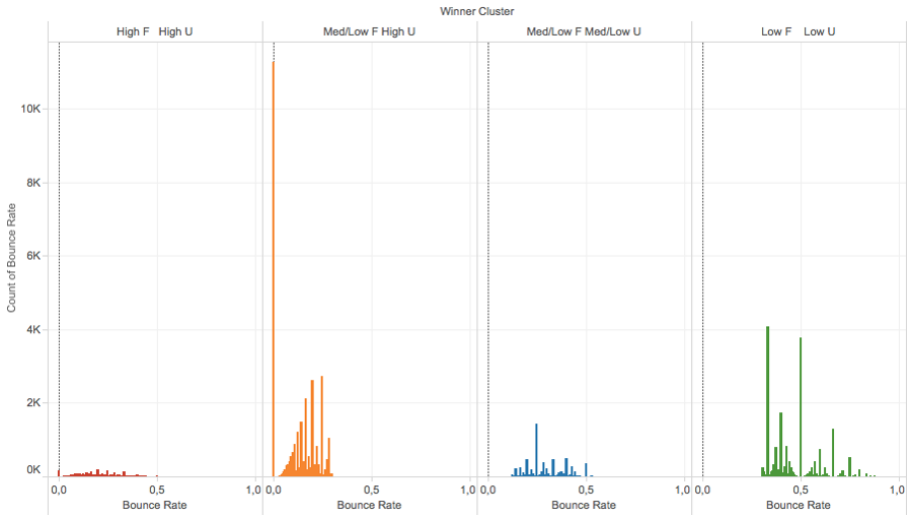

Figura 4.17: Istogramma relativo alla distribuzione del bounce rate rispetto ai quattro profili.

Questo secondo istogramma mostra la distribuzione del bounce rate rispetto ai quattro clusters. Anche in questo caso, trattandosi di una metrica molto discriminatoria, è facile intuire come nel gruppo 1 i valori siano tendenzialmente bassi (sotto il 50%), mentre nel gruppo 4, la maggior parte dei valori tende verso il 100%.

Winner Cluster	Min. AR	Max. AR	Min. AVG	Max. AVG	Min. Visit Per Day	Max. Visit Per Day	Min. AVG NumPages PerVisit	Max. AVG NumPages PerVisit	Min. AVG Time On Site Per V..	Max. AVG Time On Site Per V..	Min. Bounce Rate	Max. Bounce Rate
High F High U	0.304	1.000	1.000	4.500	1.000	4.600	1.056	24.098	2.921	1,032.067	0.000	0.529
Med/Low F High U	0.027	0.379	1.000	67.000	1.000	2.381	1.048	26.000	0.456	1,842.000	0.000	0.304
Med/Low F Med/Low U	0.027	0.479	1.000	70.000	1.857	4.750	1.063	17.000	0.278	1,692.542	0.120	0.529
Low F Low U	0.027	0.479	1.000	69.000	1.000	2.619	1.019	15.227	0.074	1,122.500	0.308	0.981

Figura 4.18: Tabella dei valori massimi e minimi di ogni metrica per ogni cluster

La tabella riporta i valori massimi e minimi di ogni metrica per ogni cluster, per poter studiare meglio la distribuzione dei valori di ogni metrica.

Ecco, quindi, sulla base dei dati analizzati, un'interpretazione dei quattro profili individuati:

- **gruppo HIGH F HIGH U** (15% del dataset) All'interno di questo cluster sono raccolti tutti gli utenti che possiedono un profilo alto sia dal punto di vista della frequenza che del tasso di utilizzo;
- **gruppo MEDIUM F HIGH U** (46% del dataset) E' il cluster più popolato, quasi la metà di tutti gli utenti analizzati. Raccoglie tutti i jobseeker che presentano un profilo medio per la frequenza di accesso, e un profilo alto per l'utilizzo del portale;
- **gruppo MEDIUM F MEDIUM U** (11% del dataset) In questo gruppo troviamo utenti che presentano un profilo medio/basso per frequenza di accesso, e un profilo medio per l'utilizzo;
- **gruppo LOW F LOW U** (28% del dataset) Raccoglie tutti gli utenti caratterizzati da un profilo basso in tutto.

Da questa analisi è stato possibile individuare quattro profili distinti, caratterizzati da diversi comportamenti di navigazione all'interno del sito di Jobrapido. Questo tipo di informazione risulta essere molto utile sia a fine descrittivo che per poter procedere con le analisi successive. Per scopo descrittivo si intende che, tramite questa analisi, Jobrapido ha una prima visione sulla tipologia dei suoi utenti (analizzati in Italia tra il gennaio e il marzo 2015).
Grazie a questo tipo di analisi, si è in grado di assegnare un'etichetta a ogni singolo jobseeker, in modo tale da poter effettuare delle analisi di classificazione supervisionata, e capire quindi le possibili differenze che probabilmente portano un utente a far parte di un gruppo piuttosto che di un altro.

Lo scopo principale sarà quello di capire tali differenze, come poter sfruttare determinati attributi per far leva sui jobseekers e tentare di spostare gli utenti nei gruppi sempre più alti e "remunerativi".

Capitolo 5

Analisi di Classificazione

5.1 Processo di classificazione

In questo capitolo verrà affrontata la seconda fase analitica legata alla **modellazione degli utenti** attraverso gli algoritmi di classificazione, partendo dai risultati ottenuti nella fase di clustering.
Lo scopo di questa analisi è quella di costruire dei modelli di previsione che possano classificare gli utenti in base al loro comportamento, cercando di determinare il profilo cui appartengono. Al termine della prima analisi, come illustrato nel capitolo precedente, si è riusciti ad associare a ogni jobseeker un determinato profilo. A partire da questo, l'idea di base è stata quella di indagare sulle possibili differenze tra i vari cluster, per capire se esistevano delle motivazioni che avevano portato gli utenti ad appartenere a un gruppo piuttosto che a un altro, come il tipo di esperienza avuto, o a seguito degli annunci comparsi o delle keyword utilizzate.
Quindi l'approccio iniziale è stato quello di pensare a una serie di metriche che avrebbero permesso di rispondere a queste domande. Successivamente le stesse saranno utilizzate negli algoritmi di classificazione con lo scopo di costruire delle regole predittive in grado di classificare gli utenti e scoprire eventuali differenze tra i vari cluster. Nei successivi paragrafi verranno descritte tutte le metriche costruite e utilizzate nell'analisi di classificazione. Nella parte finale verranno mostrati i modelli individuati e le interpretazioni a seguito dei risultati ottenuti.

5.2 Costruzione delle metriche

In questo paragrafo vengono illustrate le **metriche** create per descrivere l'esperienza sul portale e il comportamento assunto dagli utenti, che saranno poi utilizzate come input per l'algoritmo di classificazione.
Per ciascuna di esse verrà mostrata la formula che ne permette il calcolo, riportando i vari attributi utilizzati per la loro costruzione.
In alcuni casi, come intervallo temporale, è stata considerata solo la prima settimana a

seguito della data di conferma di ogni utente, perché da alcuni studi condotti internamente all'azienda è emerso che alcune differenze sono visibili nei primi momenti di vita dell'utente e non nel lungo periodo.
Viceversa, per altre metriche è stato considerato l'intero periodo di vita, in quanto era necessario studiare tutto il loro percorso e tutte le loro azioni. In ogni caso, per ogni metrica, viene specificato il periodo di tempo considerato. Le metriche sono le seguenti:

- **Distinct Jobsite Variety Daily**

$$\frac{(SUM(Distinct\ Count\ FK\_CUSTOMER))}{(Distinct\ Count\ FK\_DATE)}$$

  Questa metrica descrive il numero distinto di jobsite (quindi il sito che riporta gli annunci di lavoro) visto da ogni utente in media al giorno. E' stato preso in considerazione come intervallo temporale una settimana dalla data di conferma del jobseeker esaminato.

- **Distinct Jobsite Variety In Period**

$$\frac{(SUM(Distinct\ Count\ FK\_CUSTOMER\_InTot))}{(Distinct\ Count\ FK\_DATE)}$$

  Si intende il numero distinto di jobsite che ogni utente ha visto nel periodo di tempo analizzato. E' stato preso in considerazione come intervallo temporale una settimana dalla data di conferma del jobseeker esaminato.

- **AVG Distinct Jobsite Per Page**

$$\frac{(SUM(Distinct\ Count\ FK\_CUSTOMER))}{(Distinct\ Count\ PageViewUUID)}$$

  Si intende il numero distinto di jobsite che ogni utente ha visto in media per pagina. E' stato preso in considerazione come intervallo temporale una settimana dalla data di conferma del jobseeker esaminato.

- **AVG Distinct Jobsite Per Visit**

$$\frac{(SUM(Distinct\ Count\ FK\_CUSTOMER))}{(Distinct\ Count\ VisitUUID)}$$

  Si intende il numero distinto di jobsite che ogni utente ha visto in media per visita. E' stato preso in considerazione come intervallo temporale una settimana dalla data di conferma del jobseeker esaminato.

- **AVG Distinct Jobsite 1 Page Daily**

$$\frac{(SUM(Distinct\ Count\ FK\_CUSTOMER))}{(Distinct\ Count\ FK\_DATE)}$$

Si intende il numero distinto di jobsite che ogni utente ha visto in media giornalmente nella prima pagina. E' stato preso in considerazione come intervallo temporale una settimana dalla data di conferma del jobseeker esaminato.

- **Distinct Advert Variety Daily**

$$\frac{(SUM(Distinct\ Count\ FK\_ADVERT))}{(Distinct\ Count\ FK\_DATE)}$$

Questa metrica descrive il numero distinto di advert (quindi l'annuncio di lavoro, e non il sito che riporta tale annuncio) che ogni utente vede in media al giorno. E' stato preso in considerazione come intervallo temporale una settimana dalla data di conferma del jobseeker esaminato.

- **Distinct Advert Variety In Period**

$$\frac{(SUM(Distinct\ Count\ FK\_CUSTOMER))}{(Distinct\ Count\ FK\_DATE)}$$

Si intende il numero distinto di advert che ogni utente ha visto nel periodo di tempo analizzato. E' stato preso in considerazione come intervallo temporale una settimana dalla data di conferma del jobseeker esaminato.

- **AVG Distinct Advert Per Page**

$$\frac{(SUM(Distinct\ Count\ FK\_ADVERT))}{(Distinct\ Count\ PageViewUUID)}$$

Si intende il numero distinto di advert che ogni utente ha visto in media per pagina. E' stato preso in considerazione come intervallo temporale una settimana dalla data di conferma del jobseeker esaminato.

- **AVG Distinct Advert Per Visit**

$$\frac{(SUM(Distinct\ Count\ FK\_ADVERT))}{(Distinct\ Count\ VisitUUID)}$$

Si intende il numero distinto di advert che ogni utente ha visto in media per visita. E' stato preso in considerazione come intervallo temporale una settimana dalla data di conferma del jobseeker esaminato.

- **AVG Distinct Advert 1 Page Daily**

$$\frac{(SUM(Distinct\ Count\ FK\_ADVERT))}{(Distinct\ Count\ FK\_DATE)}$$

Si intende il numero distinto di advert che ogni utente ha visto in media giornalmente nella prima pagina. E' stato preso in considerazione come intervallo temporale una settimana dalla data di conferma del jobseeker esaminato.

- **Advert per Distinct Jobsite**

$$\frac{(SUM(Distinct\ Count\ FK\_ADVERT))}{(DistinctCount\ FK\_CUSTOMER)}$$

Questa metrica esprime il rapporto tra il numero di annunci distinti e il numero di jobsite distinti, per ogni utente. E' stato preso in considerazione come intervallo temporale una settimana dalla data di conferma del jobseeker esaminato.

- **Label Device**
  Rappresenta un'etichetta assegnata in relazione al device più utilizzato dal jobseeker. E' stato preso in considerazione come intervallo temporale l'intera vita del jobseeker. Le etichette sono:
  - *Mobile_Only* = l'utente ha utilizzato solo il dispositivo mobile per visitare il portale.
  - *Desktop_Only* = l'utente ha utilizzato solo il pc per visitare il portale.
  - *Tablet_Only* = l'utente ha utilizzato solo il tablet per visitare il portale.
  - *Mobile_Cross* = l'utente ha utilizzato più dispositivi, ma la maggior parte delle visite sono state effettuate da un dispositivo mobile.
  - *Desktop_Cross* = l'utente ha utilizzato più dispositivi, ma la maggior parte delle visite sono state effettuate dal pc.
  - *Tablet_Cross* = l'utente ha utilizzato più dispositivi, ma la maggior parte delle visite sono state effettuate da tablet.

- **Label Time**
  Rappresenta un'etichetta assegnata in relazione al periodo della giornata in cui il jobseeker ha effettuato la maggior parte delle visite. E' stato preso in considerazione come intervallo temporale l'intera vita del jobseeker. Le etichette sono:
  - *Most_Mattina* = dalle 7 a.m. alle 12 a.m.
  - *Most_Pomeriggio* = dal 1 p.m. alle 6 p.m.
  - *Most_Sera* = dalle 7 p.m. alle 12 p.m.
  - *Most_Notte* = dal 1 a.m. alle 6 a.m.

- **Test**
  Etichetta (yes/no) che indica se un utente ha partecipato ad almeno un test oppure no. E' stato preso in considerazione come intervallo temporale l'intera vita del jobseeker.

- **% Sponsored**
$$\frac{(SUM(is\_Sponsored))}{(Count\ FK\_ADVERT)}$$
  Questa metrica misura il numero di annunci sponsorizzati visti da ogni jobseeker in percentuale. E' stato preso in considerazione come intervallo temporale l'intera vita del jobseeker.

- **% Traffic Source Mix**
$$\frac{(Distinct\ Count\ VisitUUID\ (con\ source\ != JobAllert))}{(Distinct\ Count\ VisitUUID\ (con\ JobAllert))}$$
  Questa metrica esprime in percentuale, per ogni utente, il mix relativo alla varietà delle sorgenti da cui l'utente è arrivato sul sito, escludendo le JobAllert. E' stato preso in considerazione come intervallo temporale l'intera vita del jobseeker

- **% Words**
$$\frac{(SUM(ricerche\_words))}{Count\ VisitUUID}$$
  E' una percentuale relativa al numero di parole chiave nel campo word utilizzate da ogni utente, rispetto al numero di visite. Un valore pari al 100% indica che su 10 visite l'utente ha effettuato 10 ricerche riempendo sempre e solo il campo word. E' stato preso in considerazione come intervallo temporale l'intera vita del jobseeker

- **% Locations**
$$\frac{(SUM(ricerche\_location))}{Count\ VisitUUID}$$
  Identica alla precedente, questa metrica esprime una percentuale relativa al numero di parole chiave nel campo location utilizzate da ogni utente, rispetto al numero di visite. Un valore pari al 100% indica che su 10 visite l'utente ha effettuato 10 ricerche riempendo sempre e solo il campo location. E' stato preso in considerazione come intervallo temporale l'intera vita del jobseeker

- **% Both WordsLocations**

$$\frac{(SUM(ricerche\_both))}{Count\ VisitUUID}$$

Identico alle due metriche precedenti, in questo caso si considerano le ricerche in cui sono stati riempiti sia il campo word che il campo location. E' stato preso in considerazione come intervallo temporale l'intera vita del jobseeker

- **AVG Statement**

$$\frac{(SUM(SEARCH\_NUM))}{Distinct\ Count\ VisitUUID}$$

Questa metrica esprime la media delle ricerche salvate da ogni utente, rapportandola al numero di visite. Per esempio un valore pari a 5 significherà che l'utente in media per visita ha 5 ricerche salvate. Il numero massimo è 10, in quanto Jobrapido permette agli utenti confermati di salvare fino a 10 ricerche, in modo tale da consigliare all'utente annunci di lavoro che corrispondono alle ricerche salvate. E' stato preso in considerazione come intervallo temporale l'intera vita del jobseeker.

- **% Mobile Visits**

$$\frac{(SUM(Visite\ fatte\ da\ Mobile))}{Distinct\ Count\ VisitUUID}$$

Percentuale che esprime il numero di visite effettuate da Mobile, rispetto al numero totale di visite effettuate. E' stato preso in considerazione come intervallo temporale l'intera vita del jobseeker

- **AVG Quality Score**

$$\frac{(SUM(quality\ Score))}{Count\ VisitUUID}$$

Questa metrica esprime una media legata alla qualità dei jobsite visti dall'utente. Il calcolo è stato ottenuto utilizzando una dataset interno a Jobrapido che, secondo determinate caratteristiche, assegnava un punteggio di qualità ai siti che ospita sul portale. Questo ci ha permesso di effettuare una media dei punteggi per ogni utente, per capire in media la qualità dei siti che ogni utente ha visionato. E' stato preso in considerazione come intervallo temporale l'intera vita del jobseeker

- **HR**

$$\frac{(Distinct\ Count\ (code\_Customer\_HR))}{(Distinct\ Count(code\_Customer))}$$

E' un valore che esprime la percentuale di annunci HR (cioè l'annuncio che porta direttamente al sito aziendale) visti da un utente. E' stato preso in considerazione come intervallo temporale l'intera vita del jobseeker

- **% Subscription source Type**
  E' un attributo che esprime il tipo di sorgente che ha portato l'utente su Jobrapido, dove successivamente si è sottoscritto. In questo caso la sorgente può essere di tipo paid o free.

- **Gender**
  Attributo che indica se l'utente è maschio o femmina. Jobrapido non possiede questo tipo di informazione, così è stato eseguito un piccolo esperimento, in cui si è deciso di provare a derivare questo tipo di informazione dall'indirizzo email. E' stato implementato un algoritmo in Java per determinare questo tipo di informazione, incrociando gli indirizzi email con un dataset contenente l'elenco dei nomi italiani. Il risultato è stato parzialmente soddisfacente, ma incompleto, e quindi non è stato molto utile al fine della classificazione.

- **Num_Distinct_Words/NumDistinctSearches**

$$\frac{Number\ Distinct\ Words}{Count(Distinct\ Searches)}$$

  Questa metrica esprime la varietà delle ricerche effettuate da ogni utente. Infatti viene effettuato il rapporto tra il numero di parole distinte utilizzate nel periodo di tempo analizzato e il numero di ricerche. Perciò se il valore è pari al 60%, vuol dire che l'utente su 10 ricerche ha utilizzato sei parole distinte. Ovviamente questa distinzione non è stata effettuata su base semantica, ma solo in base all'esatta parola utilizzata. Il tutto è stato realizzato tramite un algoritmo scritto in Java. Il periodo di tempo preso in esame è pari all'intera vita del jobseeker.

- **% Returning Traffic Mix Community**

$$\frac{(Distinct\ Count\ VisitUUID(codeType\ =\ Community))}{Distinct\ Count\ VisitUUID}$$

  Questa metrica esprime in percentuale, per ogni utente, il numero di volte che è tornato su Jobrapido da community (quindi attraverso il click fatto sulla joballert inviata dal portale stesso), rispetto al totale delle visite effettuate. In questo caso è stato preso in considerazione come intervallo temporale l'intera vita del jobseeker.

- **% Returning Traffic Mix Paid**

$$\frac{(Distinct\ Count\ VisitUUID(codeType\ =\ Paid))}{Distinct\ Count\ VisitUUID}$$

  Questa metrica esprime in percentuale, per ogni utente, il numero di volte che è tornato su Jobrapido da una fonte a pagamento (per esempio tramite AdWords, Bing ecc..), rispetto al totale delle visite effettuate. In questo caso è stato preso in considerazione come intervallo temporale l'intera vita del jobseeker.

- **% Returning Traffic Mix Brand**

$$\frac{(Distinct\ Count\ VisitUUID(codeType\ =\ Brand))}{Distinct\ Count\ VisitUUID}$$

Questa metrica esprime in percentuale, per ogni utente, il numero di volte che è tornato su Jobrapido da una fonte brand, rispetto al totale delle visite effettuate. In questo caso è stato preso in considerazione come intervallo temporale l'intera vita del jobseeker.

- **% Returning Traffic Mix Free**

$$\frac{(Distinct\ Count\ VisitUUID(codeType\ =\ Free))}{Distinct\ Count\ VisitUUID}$$

Questa metrica esprime in percentuale, per ogni utente, il numero di volte che è tornato su Jobrapido da una fonte free (per esempio un risultato organico di un motore di ricerca), rispetto al totale delle visite effettuate. In questo caso è stato preso in considerazione come intervallo temporale l'intera vita del jobseeker.

- **Subscription Source Adwords**
  Attributo (yes/no) che indica se l'utente esaminato si è sottoscritto a Jobrapido provenendo da Adwords.

- **Subscription Source Yahoo**
  Attributo (yes/no) che indica se l'utente esaminato si è sottoscritto a Jobrapido provenendo da Yahoo.

- **Subscription Source Criteo**
  Attributo (yes/no) che indica se l'utente esaminato si è sottoscritto a Jobrapido provenendo da Criteo.

- **Subscription Source Others**
  Attributo (yes/no) che indica se l'utente esaminato si è sottoscritto a Jobrapido provenendo da altre fonti.

- **% ClickText / TOT Click**

$$\frac{Count\ Click\ Text}{Count\ Tot\ Click}$$

Questa metrica esprime il rapporto tra il numero dei click su Ad testuali rispetto al numero di click totale. Questo valore viene espresso in percentuale. Quindi se un utente presenta un valore pari al 50% vuole dire che metà dei click che ha effettuato sono stati su Ad testuali. E' stato preso in considerazione come intervallo temporale l'intera vita del jobseeker.

- **% Page 0 Results**

$$\frac{(Count\ Page\ with\ 0\ Results\ Visited)}{Count\ Total\ Page\ Visited}$$

  Questa metrica esprime il numero di pagine che non presentavano alcun tipo di risultato, rispetto al numero di pagine totali. Il valore è espresso in percentuale. E' stato preso in considerazione come intervallo temporale l'intera vita del jobseeker.

- **Indirizzo IP**
  Questa metrica è stata calcolata incrociando gli indirizzi IP degli utenti con un dataset che conteneva le rispettive posizioni geografiche, con un dettaglio che partiva dalla città fino allo stato. E' stato scritto un programma java che ci ha permesso di incrociare questi dati e se un utente presentava più localizzazioni, si prendeva quella che compariva il maggior numero di volte. E' stato preso in considerazione come intervallo temporale l'intera vita del jobseeker.

- **Target Mail Delivery**
  Attributo (y/n) che indica se l'utente ha ricevuto almeno una volta una target mail. La target mail non è altro che un email personalizzata per l'utente stesso, diversa dalla Joballert, inviata solo quando si riscontra un match con un possibile annuncio di lavoro.

- **Num Target Mail**
  Numero di target mail ricevute. E' stato preso in considerazione come intervallo temporale l'intera vita del jobseeker.

- **Weekend/Feriale**
  Metrica che esprime se un'utente ha effettuato la maggior parte delle visite in giorni feriali o in giorni festivi. E' stato preso in considerazione come intervallo temporale l'intera vita del jobseeker.

5.3 Alberi di decisione

Una volta che le metriche relative al comportamento e all'esperienza dell'utente sono state costruite e calcolate, si è passati alla fase operativa, eseguendo le **analisi di classificazione**. Lo scopo principale di queste analisi è quello di trovare una serie di regole predittive, che permettano di classificare, con una buona accuratezza gli utenti, in base ai valori assunti nelle varie metriche. Questo ci ha permesso di individuare le differenze tra i vari profili e capire le possibili motivazioni che hanno portato un utente ad appartenere a un gruppo piuttosto che a un altro. La costruzione dell'albero di decisione è una delle tecniche più largamente utilizzata nelle analisi di classificazioni, poiché oltre ad essere una delle più affidabili, offre un modello interpretabile anche per utenti non esperti di Data Mining.

Per effettuare questo tipo di analisi è stato scelto il software Weka, utilizzando l'algoritmo J48, vale a dire la versione open source scritta in Java dell'algoritmo C4.5 adottato in Weka. L'algoritmo implementa una classificazione basata sul concetto di entropia.

L'algoritmo presenta diversi parametri, che devono essere impostati a seconda delle proprie esigenze. L'analisi è stata svolta eseguendo tantissime prove verificando diverse combinazioni delle metriche sopra descritte, cercando di trovare gli attributi che classificavano, con un'elevata accuratezza, il comportamento degli utenti.
I diversi tentativi non sono stati compiuti in modo casuale, ma con la logica di creare gruppi di attributi che potevano descrivere realmente un determinato tipo di comportamento di un utente.
Ciò è stato determinato sia dal proprio intuito nello scegliere la combinazione di attributi migliori, sia basandosi talvolta sull'esperienza di chi lavora con questo tipo di dati ogni giorno.

Un esempio può essere relativo alle metriche *%Mobile_Visits* e *%Clicktext/ClickTot*. In questo caso, le due metriche non sono mai state utilizzate entrambe per determinare delle nuove regole di classificazione, in quanto i click sulle Ad. Testuali non possono essere effettuati in un contesto mobile, perché queste sono presenti e visibili all'utente solo se lo stesso è collegato tramite un PC, e quindi visibili solo in un contesto Desktop.

Inoltre, per poter trovare la migliore combinazione possibile, e quindi con l'accuracy migliore, sono stati impostati anche i diversi parametri dell'algoritmo, a seconda del contesto e degli scopi prefissati. Per una migliore comprensione, si riportano di seguito i parametri che potevano essere settati per poter utilizzare l'algoritmo J48:

- binarySplits: se impostato a true, viene forzato lo split binario di attributi nominali;
- confidenceFactor: è usato per il pruning (letteralmente "potatura"): serve per decidere se tagliare o meno un determinato sottoalbero; a valori maggiori del condence-factor corrisponde un albero più complesso; In questo caso questo attributo si riferisce ad una fase di post-pruning, quindi la "potatura" avviene solo dopo che l'algoritmo ha generato il suo risultato, per poter semplificare l'albero creato;
- Debug: se impostato a true, il classificatore può dare informazioni addizionali in consolle;
- minNumObj: il numero minimo di istanze per foglia, quindi il numero minimo di oggetti che deve contenere un nodo;
- numFolds: determina la quantità di dati usati per ridurre l'errore dovuto al pruning. Una partizione (fold) è usata per il pruning, il resto per sviluppare l'albero;

- reducedErrorPruning: se tale campo viene settato, si effettua un determinato tipo di potatura invece del pruning di default dell'algoritmo C.4.5;
- saveInstanceData: per salvare il training set per la visualizzazione;
- Seed: il seme usato per mescolare i dati quando il "reduced-error pruning" viene settato;
- subtreeRaising: permette di sostituire un ramo con un sottoramo, piuttosto che con un nodo unico;
- unpruned: se impostato a true non viene effettuato il post-pruning.

Per effettuare queste analisi sono stati settati diversi attributi. Per evitare il problema dell'overfitting si è fatto ricorso alla tecnica del post-pruning, quindi si è valutato l'andamento dell'accuracy al variale del parametro **Confidence Factor**. Si è inoltre forzato il Binary Split per attributi categorici. Il parametro **minNumObj** variava a seconda delle esigenze, per cercare di riscontrare il modello con la migliore accuratezza possibile.

L'analisi è stata effettuata tra i due gruppi estremi (High_F High_U e Low_F Low_U), in modo tale da riuscire a riscontrare le differenze più marcate e importanti, in quanto gli utenti appartenenti a questi due profili risultano essere molto distinti tra loro.

Il dataset è composto da 22206 utenti, di cui 12462 appartengono al profilo Low_F Low_U (circa il 56%), mentre 9744 appartengono al profilo High_F High_U (circa il 44%). Per effettuare le analisi utilizzando un training set e un test set è stato usato il metodo della cross-validation.
La cross-validation è una tecnica statistica utilizzabile in presenza di una buona numerosità del campione osservato o training set. In particolare la k-fold cross-validation consiste nella suddivisione del dataset totale in k parti di uguale numerosità (si chiama anche k-fold validation) e, ad ogni passo, la parte (1/k)-esima del dataset viene ad essere il test set, mentre la restante parte costituisce il training set.
Così, per ognuna delle k parti (di solito k = 10) si allena il modello, evitando quindi problemi di overfitting.

5.4 Modelli e regole di classificazione

Dopo aver eseguito scrupolosamente le numerose prove basandosi sui criteri illustrati in precedenza, sono emersi **7 modelli distinti**, tutti caratterizzati da un elevata accuracy (78-85%), in grado di classificare gli utenti nei due profili, in base ai valori assunti nelle differenti metriche.
Questi modelli descrivono il loro comportamento e la loro esperienza sul sito, fornendo delle informazioni aggiuntive relativamente all'appartenenza a un profilo. Ognuno di

questi modelli viene riportato graficamente sotto forma di albero di decisione. Ogni ramo ha la propria etichetta che rappresenta l'attributo e il rispettivo valore.
Le foglie sono state rappresentate graficamente utilizzando un diagramma a torta che indica il numero di utenti coinvolti nel modello rispetto al totale degli utenti appartenenti a quel profilo, e viene anche riportata l'accuracy finale di quella determinata regola. I modelli rappresentati non sono completi, ma sono stati estratti solo i rami che sono in grado di classificare il maggior numero di jobseeker con la più alta accuracy.
Questa scelta è legata agli obiettivi aziendali, in quanto, in questo caso, è molto più utile fornire poche informazioni ma solide e che permettano di classificare il più alto numero di utenti possibili. Alcuni rami che erano in grado di classificare 10 o 100 jobseeker non sono stati riportati anche se presentavano un accuracy molto alta, appunto perché a livello aziendale è molto più utile conoscere quelle regole che descrivono il comportamento di molti più utenti.

Per far capire meglio questo tipo di ragionamento, si riporta un'immagine che rappresenta il classico output ottenuto con l'algoritmo J48 di Weka.

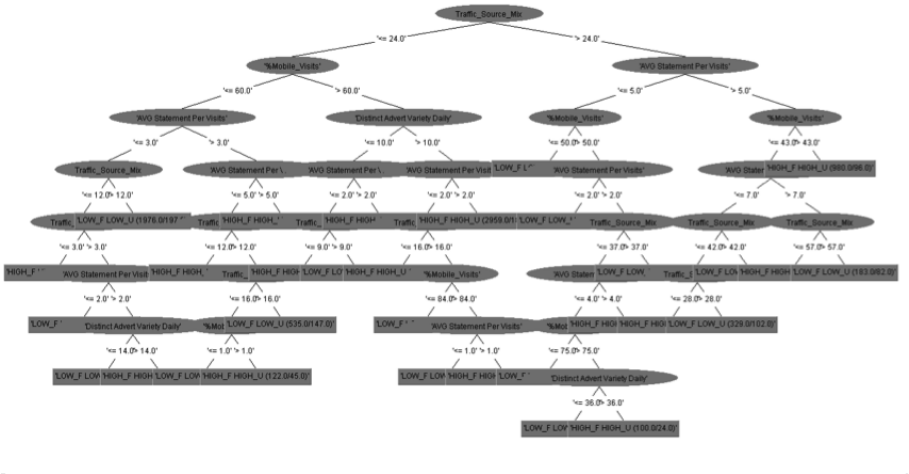

Figura 5.1: Esempio output Weka

Questa immagine rappresenta un modello completo, ottenuto dopo aver eseguito l'algoritmo J48, utilizzando 7 attributi distinti. Come si può notare, questi modelli possono risultare di difficile interpretazione, soprattutto a livello aziendale, per chi si occupa del management. Per questo motivo, dopo ogni esecuzione, si esaminava ogni modello ottenuto, cercando di estrarre i rami che possedevano una serie di attributi da formare delle regole di classificazione con un'alta accuracy e in grado di classificare un elevato numero di utenti.
Così, dopo un intenso lavoro legato alle analisi di classificazione, si è riusciti ad estrarre 7 modelli, che sono stati in grado di fornire informazioni molto interessanti al manage-

ment, il quale successivamente sarà in grado di prendere le decisioni migliori, in base alle informazioni fornite.
Ecco i vari modelli, con relativa interpretazione.

Modello 1

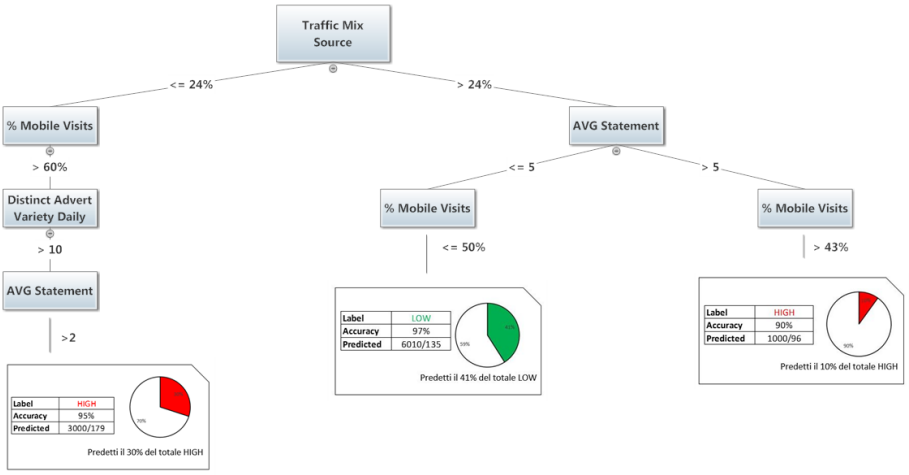

Figura 5.2: Modello di classificazione numero 1

Il primo modello è stato estratto da un albero di classificazione costruito utilizzando 6 attributi. Quattro di essi sono risultati utili alla costruzione delle regole di classificazione, mentre i restanti due non sono stati utilizzati in questo specifico caso. Come si può vedere dall'immagine, il modello permette di classificare gli utenti in entrambi i profili, sia High_F High_U, sia Low_F Low_U.

Partendo da sinistra, il **primo ramo** crea una serie di regole in grado di classificare con una accuracy pari al 95% gli utenti appartenenti al profilo più alto. Questo significa che questa regola è in grado di classificare 9,5 utenti su 10 in modo corretto. Interpretando i valori in modo puntuale, partendo dal dataset di partenza, su 3000 utenti che possedevano le caratteristiche espresse dagli attributi appartenenti a quel ramo, la predizione è stata errata solo in 179 utenti.
Quindi, se gli utenti hanno un numero di sorgenti di ritorno diverse dalla Joballert inferiore al 24%, se le loro visite Mobile sono maggiori del 60%, se giornalmente visualizzano più di 10 Advert distinti e se hanno un numero di Statement (ricerche) salvate maggiore di 2, molto probabilmente saranno utenti che utilizzano molto spesso il portale, e quindi appartenenti al profilo più alto.
Il **secondo ramo**, invece, predice gli utenti appartenenti al profilo più basso, con una accuracy pari al 97% (su 6010 utenti, vengono predetti correttamente 5875 utenti). La terza regola, invece, è in grado di predire gli utenti con una accuratezza pari al 90%, ma

come si può notare dal diagramma a torta, la percentuale di utenti utilizzati dal dataset è relativamente bassa rispetto alle altre due regole di classificazione.

Modello 2

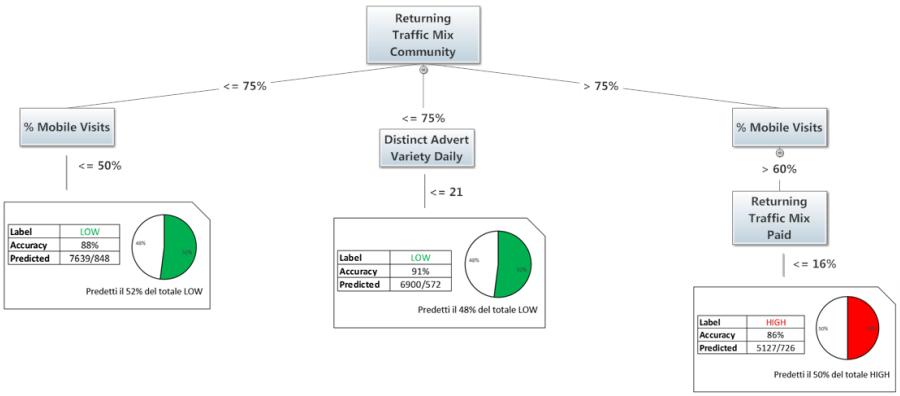

Figura 5.3: Modello di classificazione numero 2

Il secondo modello è leggermente più semplice del primo, in quanto sono state ricavate regole meno complesse, con meno attributi, ma con percentuali di copertura relative al dataset di partenza molto più alte. Rispetto al primo modello sono stati utilizzati due nuovi attributi, Returning traffic mix Community e Returning traffic mix Paid. Come si può notare, anche dal modello precedente, e come si potrà notare anche dai modelli successivi, l'attributo legato alla percentuale di visite con dispositivo Mobile è molto predittiva, sia per gli utenti High che per gli utenti Low.
Infatti, questo può essere tradotto attribuendo agli utenti più redditizi per Jobrapido un alto tasso di utilizzo del dispositivo mobile, a differenza degli utenti caratterizzati da un basso profilo. Il nodo di partenza risulta essere una buona conferma per Jobrapido, in quanto gli utenti che ritornano da Community sono gli utenti che sono arrivati su Jobrapido tramite le Joballert inviate dal portale stesso. Quindi questa è una chiara conferma di quello che già sostenevano, e che quindi il servizio di Joballert funziona bene, aiutando molto gli utenti nella loro ricerca.

Modello 3

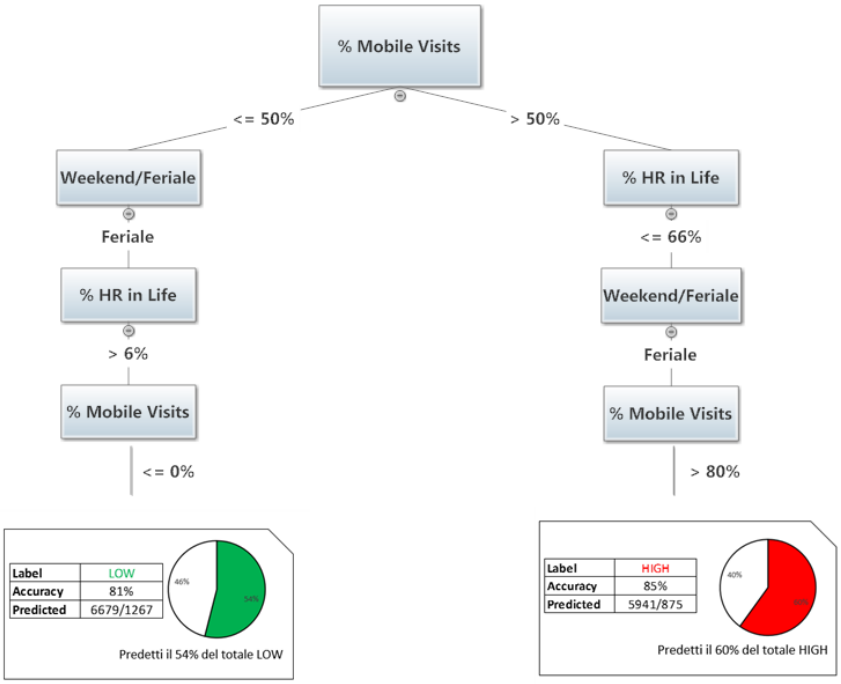

Figura 5.4: Modello di classificazione numero 3

Questo modello è molto interessante, soprattutto per l'utilizzo dell'attributo *Weekend/Feriale*, che fornisce delle informazioni curiose circa il comportamento degli utenti. In entrambe le due regole di classificazione, il modello predittivo indica i giorni feriali come quelli in cui il portale viene maggiormente utilizzato da entrambi i tipi di profilo. Ovviamente la combinazione con altri attributi permette di distinguere il tipo di utente. Entrambe le regole presentano un accuracy molto elevata, e sono in grado di predire un alto numero di utenti, comprendo più del 50% il dataset di partenza per utenti Low e per utenti High. Anche in questo modello, il numero di visite effettuate da Mobile è molto predittivo e discriminante.
Infatti, gli utenti appartenenti al profilo High_F High_U presentano un alto tasso di utilizzo del mobile per effettuare le visite su Jobrapido, a differenza degli utenti Low_F Low_U, la cui percentuale scende addirittura allo 0%.

Modello 4

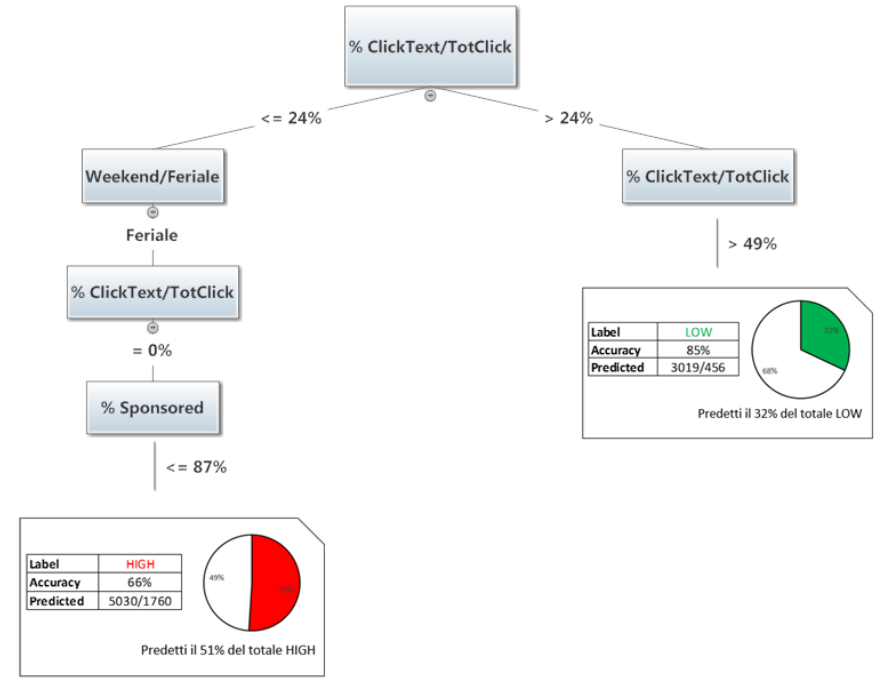

Figura 5.5: Modello di classificazione numero 4

Questo modello introduce un altro nuovo attributo molto significativo, *%clickText/TotClick*, che descrive la percentuale di click su ad. testuali sul totale dei click.
Questo attributo mostra come gli utenti appartenenti a un alto profilo sono caratterizzati dal fatto che non effettuano molti click sulle Ad. testuali, a differenza degli utenti Low. Inoltre, per gli utenti High, la percentuale di risultati sponsorizzati visualizzati è inferiore al 87%, e questo è un dato molto importante in quanto se a un utente vengono restituiti solo annunci sponsorizzati, molto probabilmente porteranno il jobseeker ad appartenere a un gruppo diverso, e quindi a tornare meno spesso sul portale.
In questo caso, il ramo che predice gli utenti High presenta un valore di accuracy che risulta essere il più basso in assoluto rispetto a tutti e 7 i modelli trovati, circa il 66%.

Modello 5

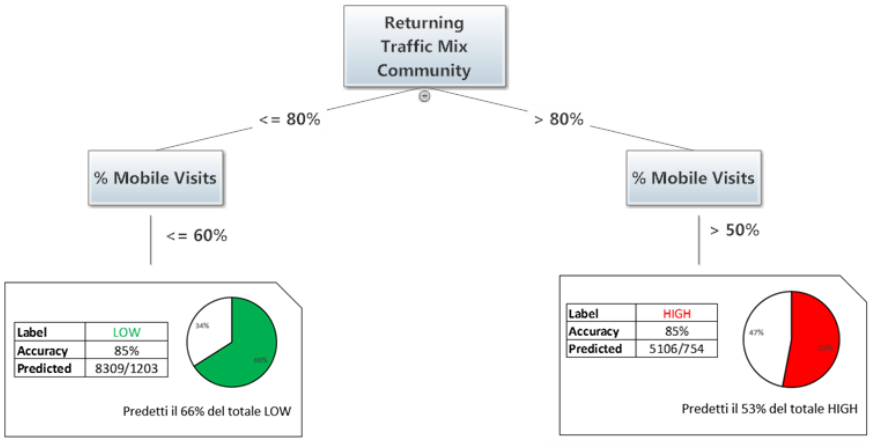

Figura 5.6: Modello di classificazione numero 5

Il modello 5 è il più semplice di tutti, ma è anche quello che riesce a predire il maggior numero di jobseeker, in quanto per il profilo Low è in grado di classificare il 66% del dataset di partenza, un numero decisamente alto.
Questo modello è un ulteriore conferma di ciò che è potuto emergere nei modelli precedenti, e cioè che gli utenti con un'alta attitudine all'utilizzo del dispositivo mobile sono solitamente gli utenti che utilizzano e tornano più spesso su Jobrapido.
L'accuratezza di entrambi i modelli è pari al 85%, con coperture che superano il 50%, quindi valori molto alti e solidi. In conclusione questo modello non aggiunge nulla di nuovo, ma è un ulteriore conferma e un rafforzamento di quanto emerso precedentemente.

Modello 6

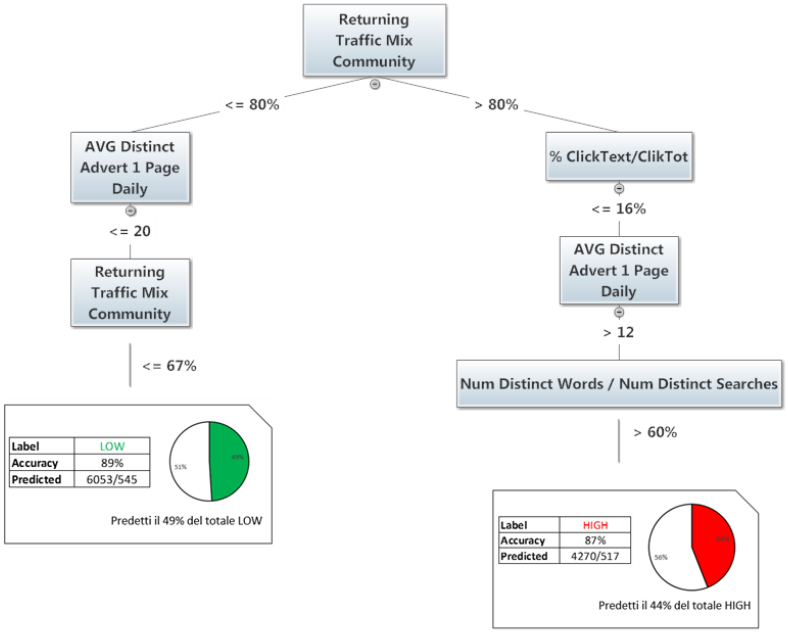

Figura 5.7: Modello di classificazione numero 6

Gli ultimi due modelli sono quasi simili e sono quelli più interessanti, caratterizzati da una varietà degli attributi utilizzati e dai valori determinati. Partendo da sinistra, tutti gli utenti che con una frequenza minore dell' 80%(che successivamente si restringe al 67%) ritornano da community (quindi tramite Joballert) e visualizzano giornalmente nella prima pagina meno di 20 advert distinti, con un' alta probabilità (accuracy pari al 89%) saranno predetti come utenti Low_F Low_U.
Nel ramo di destra, invece, tutti gli utenti che tornano da community con una frequenza maggiore dell' 80%, effettuano una quantità di click testuali minore del 16% rispetto al loro totale, giornalmente nella prima pagina visualizzano più di 12 advert distinti e hanno una varietà delle keywords maggiore del 60% (questo significa che su 10 ricerche, 6 o più sono state fatte utilizzando keywords distinte), allora saranno predetti come utenti High_F High_U con una accuratezza pari al 87%.
Questi modelli sono risultati molto interessanti per la combinazione di attributi trovata, a partire dai quali potranno essere costruite delle features operative per migliorare l?ingaggio degli utenti in Jobrapido.

Modello 7

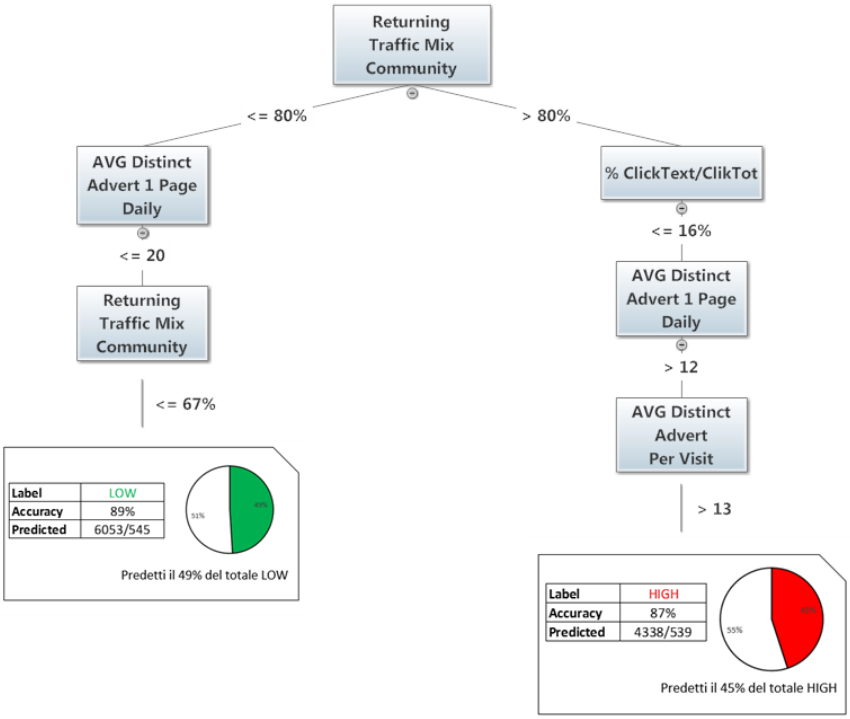

Figura 5.8: Modello di classificazione numero 7

L'ultimo modello è identico al precedente per quanto riguarda il ramo di sinistra, ma per la regola a destra, invece, presenta un cambiamento per quanto riguarda l'ultimo attributo.
In questo caso è stata inserita la media degli advert distinti visualizzati per visita. L'accuracy, rispetto al modello precedente, rimane la medesima, ma in questo caso vi è un aumento degli utenti predetti, passando al 45% del dataset di partenza.

5.5 Controllo validità delle regole di classificazione

In questa sezione viene descritto un approccio relativo al controllo della solidità delle **regole di classificazione** trovate in precedenza. I modelli che sono stati determinati attraverso l'analisi di classificazione, sono stati costruiti partendo dal dataset iniziale (country Italia) che conteneva le visite effettuate da gennaio 2015 a marzo 2015, dagli utenti sottoscritti e confermati nel mese di gennaio 2015. Quindi tutti i modelli sono stati costruiti partendo dai comportamenti di questi utenti. L'idea di base, perciò, è stata quella di prendere un nuovo gruppo di utenti, completamenti distinto dal precedente, ma che appartenessero alla medesima country, ed effettuare il processo di analisi "al contrario" per verificare la solidità di queste regole.

Il dataset esaminato contiene tutti gli utenti sottoscritti e confermati nel mese di Aprile 2015, e sono state analizzate le loro visite nei mesi di Aprile e di Maggio 2015.

Il procedimento per poter verificare la solidità dei modelli è stato quello di operare partendo dalla fine, in modo tale da verificare se le regole trovate mantenessero lo stesso livello di accuratezza anche con utenti completamente nuovi.

Quindi, riassumendo, per il dataset gennaio-marzo sono stati creati prima i profili attraverso le analisi di clustering e poi, mediante le analisi di classificazione, sono stati trovati 7 modelli predittivi. Per il dataset aprile-maggio, invece, sono state applicati subito i modelli predittivi creati con il primo dataset, e agli utenti che rispettavano le regole di classificazione è stata affidata l'etichetta in corrispondenza del profilo predetto.
Successivamente, è stata applicata la stessa analisi di clustering e si sono creati i nuovi profili. Solo a questo punto è stato possibile verificare se il profilo attribuito all'utente coincideva con il profilo determinato nella fase di clustering. In questo modo si è potuto verificare se i vari modelli mantenevano lo stesso livello di accuratezza. Il risultato è riassunto nella tabella sottostante:

Modello		DATASET GENNAIO/MARZO		DATASET APRILE/MAGGIO	
		HIGH = 9744	LOW = 12462	HIGH = 7759	LOW = 14200
		ACCURACY	% del Dataset di Partenza	ACCURACY	% del Dataset di Partenza
1	Ramo HIGH_1	90%	10%	89%	10%
	Ramo HIGH_2	95%	30%	91%	26%
	Ramo LOW	97%	41%	89%	30%
2	Ramo HIGH	86%	50%	84%	37%
	Ramo LOW_1	88%	52%	83%	30%
	Ramo LOW_2	91%	48%	84%	31%
3	Ramo HIGH	85%	60%	74%	47%
	Ramo LOW	81%	54%	70%	46%
4	Ramo HIGH	66%	51%	42%	30%
	Ramo LOW_2	85%	32%	85%	20%
5	Ramo HIGH	85%	53%	84%	33%
	Ramo LOW	85%	66%	83%	33%
6	Ramo HIGH	87%	44%	81%	43%
7	Ramo HIGH	87%	45%	81%	31%
	Ramo LOW	89%	49%	82%	30%

Figura 5.9: Tabella riassuntiva controllo regole di classificazione

Nella prima colonna vengono rappresentati i sette modelli, e per ognuno di essi i vari rami con le regole predittive per i profili High e Low. Nella seconda e nella terza colonna sono stati riportati, i valori di accuracy di ogni ramo e la percentuale di utenti predetti dal dataset di partenza, relativo al periodo di gennaio-marzo. Infine, nelle ultime due colonne sono stati descritti i valori di accuracy e le percentuali di utenti predetti dal dataset di partenza relativo al periodo aprile-maggio.

Come si può notare, il livello di accuracy rimane più o meno simile, diminuendo in alcuni casi di pochi punti percentuali. Questa è una chiara dimostrazione di come le regole predittive riscontrare siano robuste per analizzare il comportamento degli utenti

nella country Italia. Ovviamente, per poter avere risultati più precisi e per poter fare affermazioni più forti, sarà necessario proseguire con questo tipo di controlli. Ma il periodo di tempo esaminato risulta essere già molto utile per poter prendere determinate decisioni a livello operativo da parte del management.

5.6 Altri tipi di informazioni estratte dalle metriche

Molte metriche, che sono state pensate e progettate per il processo di analisi, non si sono rivelate utili al fine della creazione dei modelli predittivi. Ma, da un punto di vista aziendale invece, alcune di esse hanno apportato un contenuto informativo nuovo ed interessante per il management aziendale, fornendo alcuni tipi di informazioni curiose sul comportamento degli utenti analizzati.
Durante la realizzazione del progetto, molte informazioni che emergevano dalle svariate analisi, ma che non avevano un utilizzo pratico agli scopi del progetto, venivano comunque interpretate ed esposte mediante delle presentazioni al manager, perché avevano lo stesso un significato che poteva rivelarsi importante in altri campi e per altri tipi di scopi.
Di seguito si riportano alcune visualizzazioni grafiche create appositamente per mostrare, in modo semplice ed immediato, alcune caratteristiche relative al comportamento degli utenti, come i giorni della settimana o i periodi della giornata in cui gli utenti effettuano più visite o più click, oppure il tipo di device più utilizzato dai vari utenti(sempre relativamente al periodo temporale analizzato e agli utenti esaminati).
Il **primo istogramma** mostra la distribuzione degli indirizzi IP sul suolo italiano per regione.

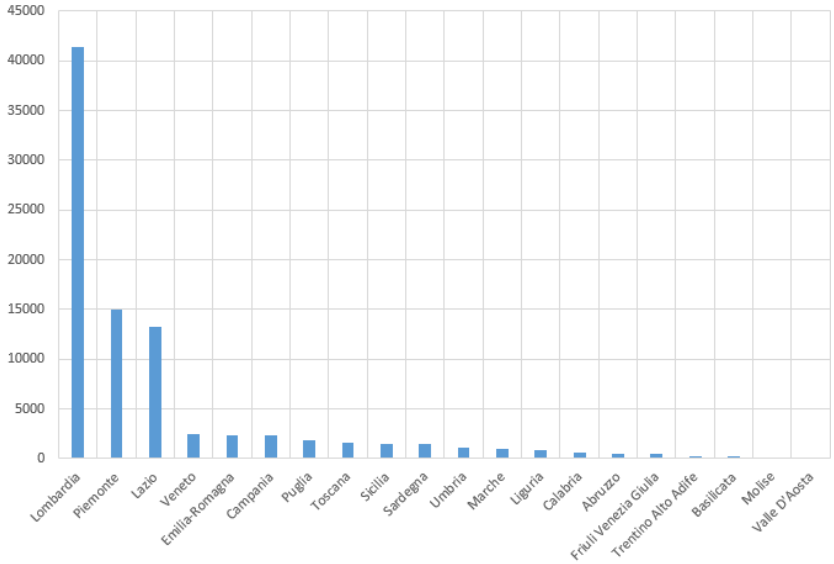

Figura 5.10: Istogramma distribuzione indirizzi IP Italia

La mappa che rappresenta lo stato italiano, invece, mostra le regioni più popolose relativamente agli utenti di Jobrapido, dove la grandezza del cerchio è direttamente proporzionale alla numerosità degli indirizzi IP. Entrambi i grafici sono relativi al pe-

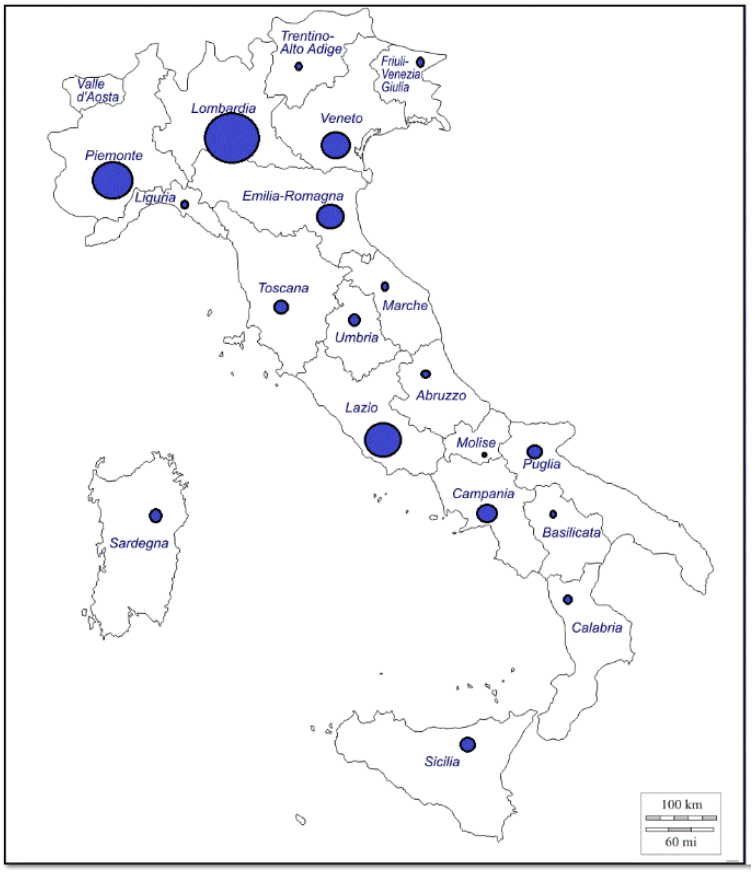

Figura 5.11: Mappa distribuzione indirizzi IP Italia

riodo di tempo tra gennaio 2015 e marzo 2015. Questi dati si sono ottenuti tramite un programma java che ho realizzato per incrociare i dati interni al portale e i dataset che contengono le coordinate geografiche relativamente agli indirizzi IP.

Il **secondo istogramma** mostra per cluster, i giorni della settimana in cui vengono effettuate il maggior numero di visite per ogni utente.Sull'asse delle Y vi è il conteggio degli utenti, in cui ogni utente ha assegnato un giorno della settimana che corrisponde al giorno in cui ha effettuato il maggior numero di visite.

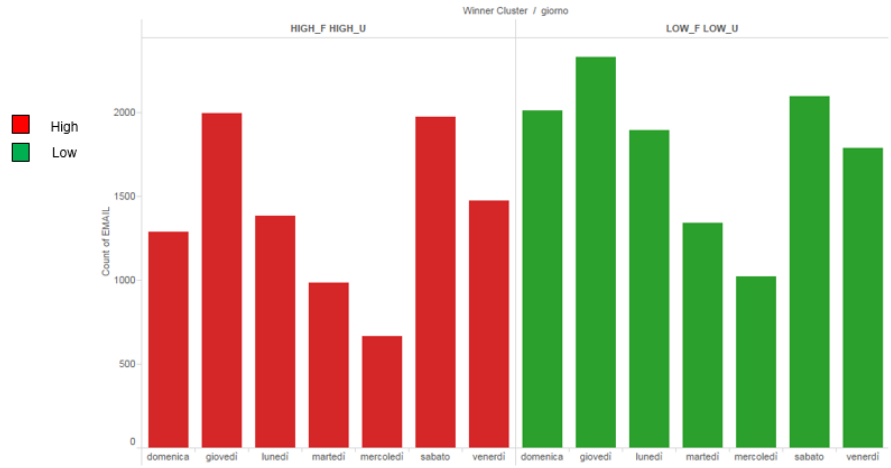

Figura 5.12: Istogramma relativo alla distribuzione delle visite nei giorni della settimana per cluster.

Come si può notare, il giovedì e il sabato risultano essere i giorni più frequenti in cui gli utenti di entrambi i profili effettuano il maggior numero di visite. La distribuzione è quasi identica per entrambi i cluster, questo quindi è stato il motivo per cui l'algoritmo di classificazione non ha ritenuto questo attributo interessante nella fase predittiva, ma è stato utile per fornire questo tipo di informazione al management relativamente al comportamento dell'utente.

Il **terzo istogramma** mostra l'andamento dei click testuali nella settimana nel cluster High_F High_U, rispettivamente nei vari momenti della giornata. È stato molto interessante scoprire come la sera (evening) era il momento della giornata in cui gli utenti effettuavano il maggior numero di click su ad.testuali. L'etichetta cross la possiedono quegli utenti che hanno effettuato lo stesso numero di click in tutti i momenti della giornata.

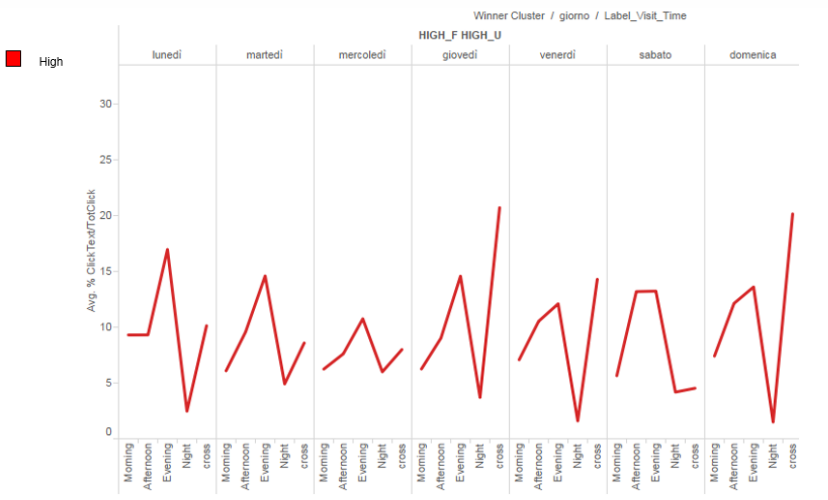

Figura 5.13: Distribuzione della percentuale dei click testuali per giorno e per momento della giornata nel cluster High

Lo stesso discorso può essere fatto nel profilo Low_F Low_U. Anche in questo caso prevale la sera (evening) come momento della giornata in cui vengono effettuati il maggior numero di click.

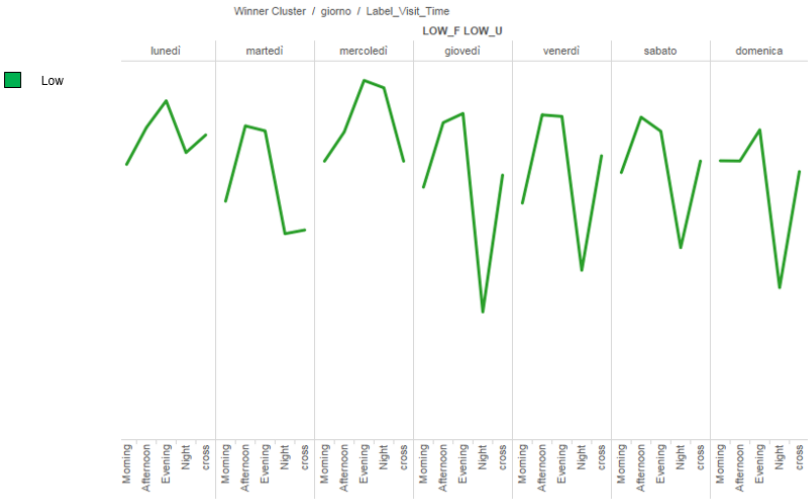

Figura 5.14: Distribuzione della percentuale dei click testuali per giorno e per momento della giornata nel cluster Low

L'istogramma seguente, invece, è relativo all'utilizzo dei vari tipi di device, diviso per cluster. Le etichette precedute da O (only) significano che l'unico device utilizzato è stato quello descritto. Se l'etichetta è preceduta da cross, invece, vuol dire che l'utente ha utilizzato più dispositivi ma quello con la maggior frequenza è quello descritto dall'etichetta.

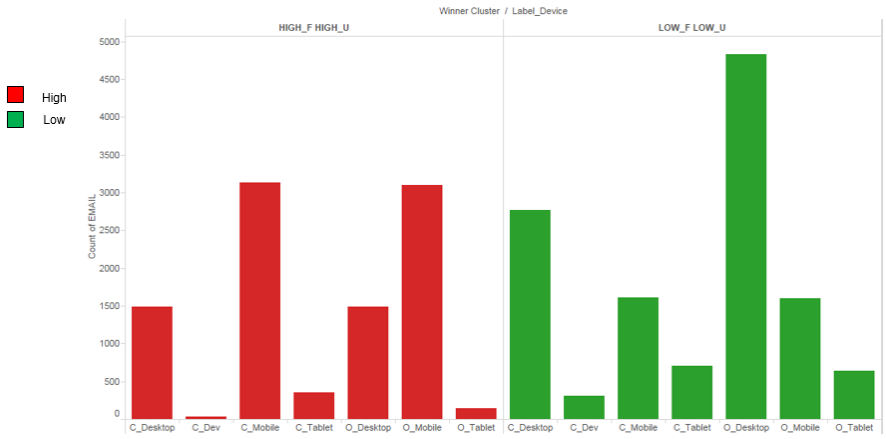

Figura 5.15: Istogramma relativo alla distribuzione dei device per cluster.

In questo caso è facile notare come per gli utenti High_F High_U il dispositivo Mobile (only o cross) risulta essere il più utilizzato, a differenza degli utenti Low_F Low_U, il cui dispositivo maggiormente utilizzato risulta essere il pc (desktop layout). Questo tipo di informazione risulta essere molto importante per il management, perché permette di avere una visione più chiara sul fatto che gli utenti che utilizzano meno il portale hanno una preferenza nell'utilizzo della versione desktop del sito e non di quella mobile.

L'ultimo istogramma, partendo da quello precedente, specifica la distribuzione dei dispositivi per cluster, scendendo ulteriormente nel dettaglio con le relative percentiali di utilizzo del cluster. Infatti ogni colore rappresenta un dispositivo e il suo relativo utilizzo (ricordiamo che *Cross* significa che l'utente ha utilizzato più dispositivi, ma per la maggior parte delle volte il device è quello specificato. *Only*, invece, significa che ha usato solo il dispositivo etichettato.)

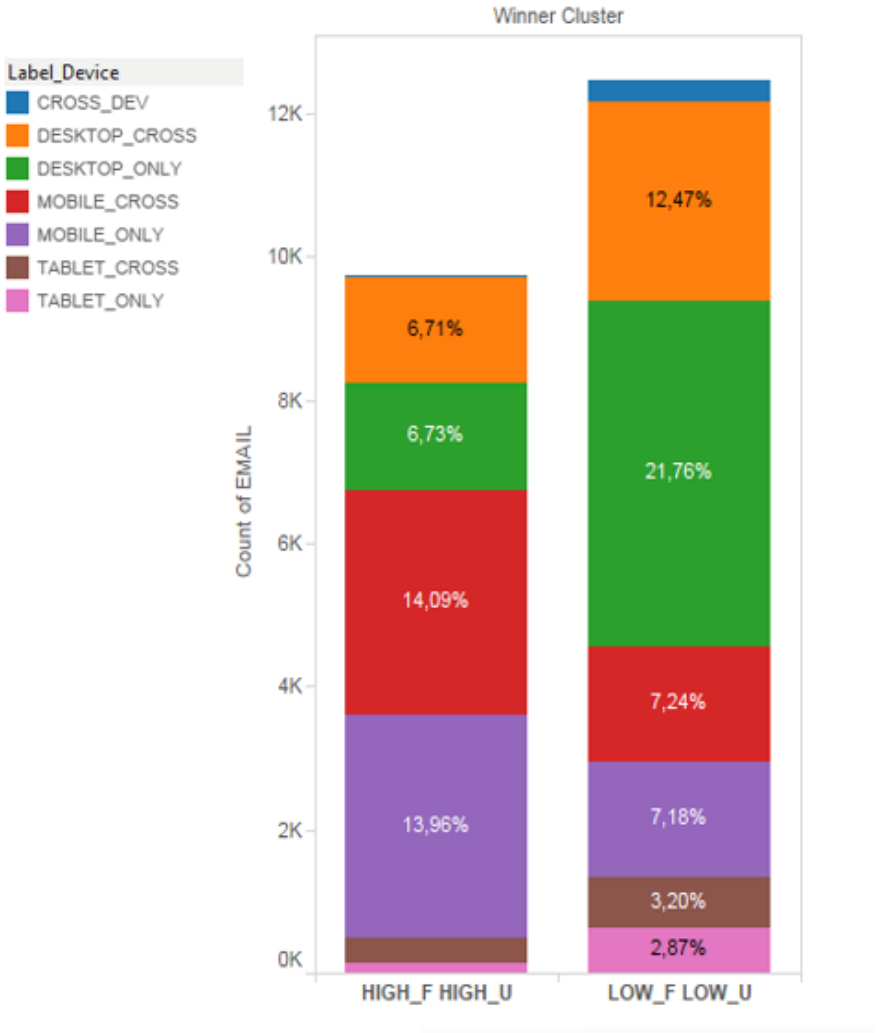

Figura 5.16: Istogramma relativo alla distribuzione dei device per cluster.

Capitolo 6

Conclusione e sviluppi futuri

L'esperienza in Jobrapido, durata 6 mesi, che ha permesso la realizzazione di questa tesi è stata proficua sia per me che per l'azienda.
Il progetto è stato un primo tentativo di individuare un metodo che permettesse di segmentare e modellare gli utenti del sito, sulla base del loro comportamento on line; inoltre, si è cercato di creare dei modelli predittivi attraverso i quali poter classificare gli utenti stessi.
Sostanzialmente, il progetto si è svolto in tre fasi.

Punto di partenza sono stati, ovviamente, i dati che quotidianamente l'azienda già raccoglieva per sé. Trattandosi di dati molto vasti, in quanto provenienti dalle interazioni di utenti appartenenti a circa 58 Paesi diversi, ed eterogenei, la prima fase è stata quella di capire quali di questi dati effettivamente potevano essere interessanti e utili alla realizzazione del nostro progetto.

Terminata la fase di data understanding, ci siamo concentrati sulla clusterizzazione degli utenti, attraverso specifiche metriche costruite ad hoc. Il risultato di questa seconda fase è stata la creazione di 4 profili tipo che hanno fornito anche nuovi tipi di informazioni all'azienda.

La terza e ultima fase è stata quella di creare dei modelli predittivi che potessero classificare gli utenti all'interno dei quattro profili stabiliti.

Per fare ciò, abbiamo costruito una serie di metriche che descrivessero, nel modo più accurato possibile, i comportamenti e le abitudini degli utenti sul portale web.
Gli obiettivi che ci eravamo prefissati sono stati tutti raggiunti e i risultati sono stati giudicati molto interessanti dall'azienda.

Gli sviluppi futuri previsti, al momento, sono di migliorare e arricchire sempre più questa metodologia di segmentazione e modellazione degli utenti, fino a sviluppare un vero

e proprio processo in real time, eseguendo giornalmente questo processo, aggiornando ogni volta i dati e ricalcolando i nuovi profili.
Tutto questo permetterebbe in tempo reale di scoprire, per esempio, se un utente ha cambiato profilo, magari passando da un profilo High a un profilo Low, in modo tale da poter intervenire tempestivamente, migliorando la sua esperienza sul sito per ricondurlo nuovamente al profilo più alto. Questo progetto deve quindi essere considerato come un punto di partenza per una futura evoluzione, soggetta a continue espansioni e ampliamenti, con ulteriori miglioramenti e ottimizzazioni soprattutto a livello processuale.

Bibliografia

[Kumar, 2006] V.Kumar, *Introduction to Data Mining*, Pearson International, 2006.

[Provost, 2012] F.Provost, *Data Science for Business - What you need to know about data mining and data-analytic thinking*, O'Reilly, 2012.

[Larose, 2012] T.Larose, *Data Mining the Web: Uncovering Patterns in Web Content, Structure, and Usage*, Wiley, 2012.

[Liu, 2012] B.Liu, *Web Data Mining - Exploring Hyperlinks, Contents, and Usage Data* Springer, 2012.

[Witten, 2005] H. Witten, *Data Mining: Practical Machine Learning Tools and Techniques*Morgan Kaufmann, 2005.

[Ghnemat, Zaghari, 2013] R. Ghnemat,R. Zaghari, *Behavioral Segmentation for Web Searching Users Using self-organizing kohonem maps*, Knowledge and Data Engineering, IEEE Transactions, 2013.

[Shang, Wang, 2009] L. Shang , L. Wang, *Web usage mining based on fuzzy clustering in identifying target group*,ISECS International Colloquium on Computing, Communication, Control, and Management, 2009.

[Hofgesang, 2014] I. Hofgesang, *Web Personalisation Through Incremental Individual Profiling and Support-based User Segmentation*, IEEE International Conference on Web Intelligence, pages 213-220, 2014.

[Saleiro, 2011] P. Saleiro, *Web sessions clustering for behavioral targeting*, 2011.

[Knime.com, 2015] Knime.com *Manuale di Knime*, https://tech.knime.org/documentation, 2015.

[r-project.org, 2015] r-project.org *Manuale di R*, http://www.r-project.org/other-docs.html, 2015.

[Frank, Hall, 2013] E.Frank,M.Hall *Weka Manual* ,http://statweb.stanford.edu, lpekelis/13_datafest_cart/WekaManual-3-7-8.pdf, 2013.

RINGRAZIAMENTI

Desidero ringraziare sentitamente il mio tutor aziendale Raffaele Serrecchia, per la grande disponibilità e cortesia dimostratemi e per tutto l'aiuto fornito durante il tirocinio. Inoltre ringrazio tutta Jobrapido per avermi accolto e per avermi dato la possibilità di fare questa esperienza lavorativa, a cominciare da Davide Conforti, Matteo Coloberti e tutti i dipendenti dell'azienda.

Ringrazio il Prof. Mirco Nanni, relatore di questa tesi, per la sua disponibilità costante e i suoi consigli ricevuti durante questo progetto.

Un ringraziamento ai miei genitori, che, con il loro sostegno morale ed economico mi hanno permesso di raggiungere questo importante traguardo.

Desidero ringraziare mio fratello che mi è sempre stato vicino e mi ha sempre sostenuto in ogni momento bello e brutto.

Infine desidero ringraziare tutti i miei amici di una vita che hanno sempre creduto in me e tutti i miei compagni di corso per essermi stati vicini durante questi anni universitari: sono stati per me non semplici compagni, ma veri amici.